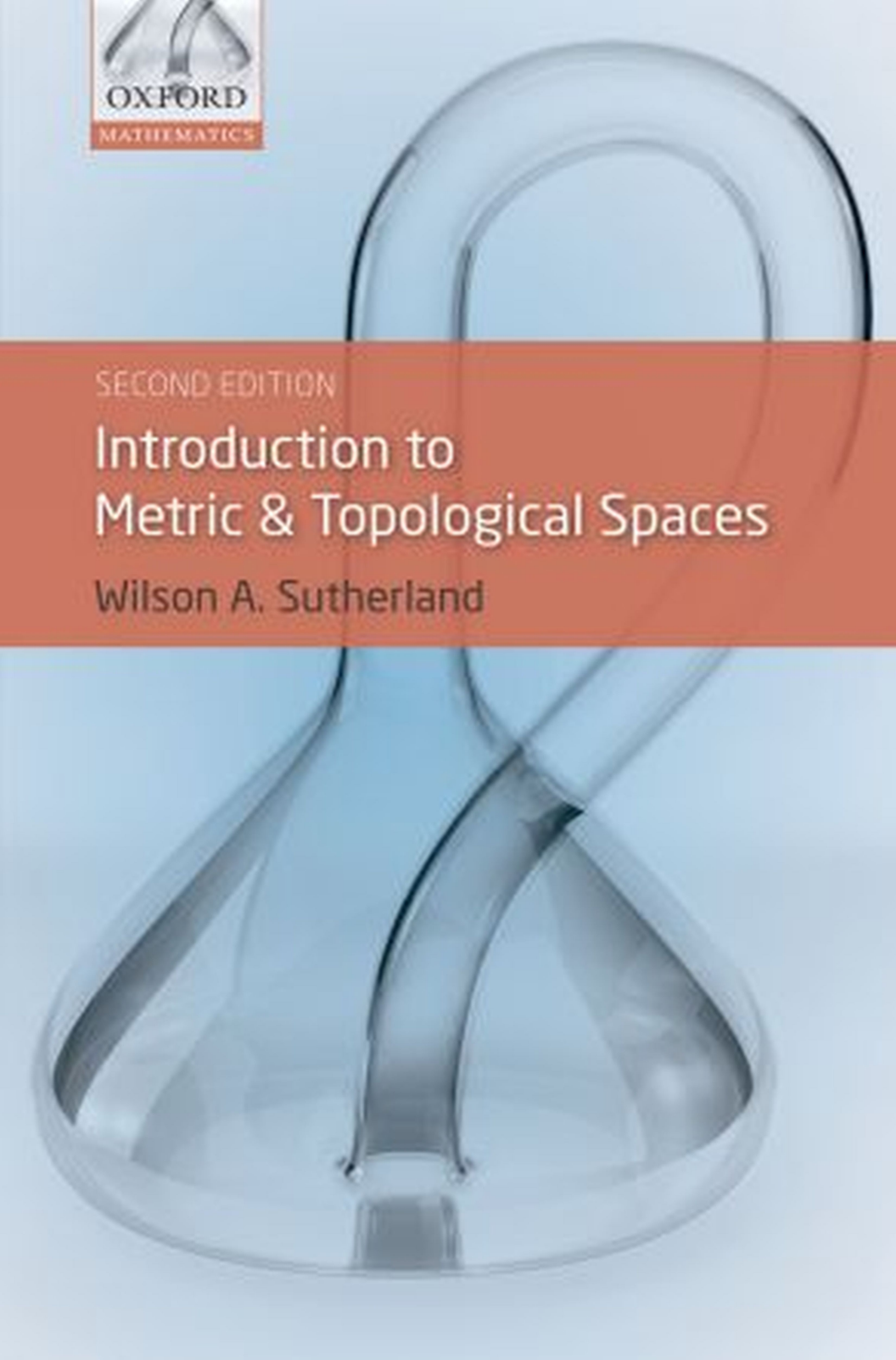
OXFORD
MATHEMATICS
SECOND EDITION
Introduction to
Metric & Topological Spaces
Wilson A. Sutherland

Introduction to Metric and Topological Spaces

Second Edition

WILSON A SUTHERLAND
Emeritus Fellow of New College, Oxford

Companion web site: www.oup.com/uk/companion/metric

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street Oxford ox2 6dp

Oxford University Press is a department of the University of Oxford
It furthers the University's objective of excellence in research scholarship
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc New York

First published 2009

British Library Cataloguing in Publication Data

Data available

Library of Congress Cataloging in Publication Data

Typeset by SPI Publisher Services Pondicherry India
Printed in Great Britain
on acid free paper by
Clays St Ives plc

ISBN 978 0 19 956807 1 (Hbk.)
978 0 19 956808 1 (Pbk.)

1 3 5 7 9 10 8 6 4 2

Preface

Preface to the second edition

One technical advance since the first edition is the possibility of having a companion web site, and I have tried to use this to the full. The address of the companion web site is www.oup.com/uk/companion/metric. Parts of the first edition have been moved there. This makes room for new material on standard surfaces, intended both to give a brief introduction to geometric topology and also to amplify the section on quotient spaces, hopefully without losing the advantage of brevity. Also, more explanations and examples have been added both to the book and on the companion web site. Accordingly the numbering in the preface to the first edition no longer applies, although the progression of ideas described there is still roughly followed. To help convey familiarity, concepts such as closure and interior are introduced first for metric spaces, and then repeated for topological spaces. I have tried to vary the accompanying examples and exercises to suit the context.

I am grateful for the opportunity to update notation and references.

A colleague who liked other aspects of the first edition complained that his students too readily looked at the answers; now as before I have a concern about students working from this book on their own, but I have moved the answers to a restricted web page.

It is a pleasure to thank anonymous referees for their thoughtful suggestions for improvements; and equally to thank two distinguished ex-students of New College for the comforting advice to change as little as possible. I hope to have steered a middle course in response to all this advice. It is also a pleasure to thank several ex-students and other friends for corrections and improvements to this edition.

It is also more than a pleasure to thank Ruth for many things, in particular her encouragement for writing a second edition.

Oxford, 2008 W.A.S.

Preface to the first edition

One of the ways in which topology has influenced other branches of mathematics in the past few decades is by putting the study of continuity and convergence into a general setting. This book introduces metric and topological spaces by describing some of that influence. The aim is to move gradually from familiar real analysis to abstract topological spaces; the main topics in the abstract setting are related back to familiar ground as far as possible. Apart from the language of metric and topological spaces, the topics discussed are compactness, connectedness, and completeness. These form part of the central core of general topology which is now used in several branches of mathematics. The emphasis is on *introduction*; the book is not comprehensive even within this central core, and algebraic and geometric topology are not mentioned at all. Since the approach is via analysis, it is hoped to add to the reader's insight on some basic theorems there (for example, it can be helpful to some students to see the Heine Borel theorem and its implications for continuous functions placed in a more general context).

The stage at which a student of mathematics should see this process of generalization, and the degree of generality he should see, are both controversial. I have tried to write a book which students can read quite soon after they have had a course on analysis of real-valued functions of one real variable, not necessarily including uniform convergence.

The first chapter reviews real numbers, sequences, and continuity for real-valued functions of one real variable. Most readers will find nothing new there, but we shall continually refer back to it. With continuity as the motivating concept, the setting is generalized to metric spaces in Chapter 2 and to topological spaces in Chapter 3. The pay-off begins in Chapter 5 with the study of compactness, and continues in later chapters on connectedness and completeness. In order to introduce uniform convergence, Chapter 8 reverts to the traditional approach for real-valued functions of a real variable before interpreting this as convergence in the sup metric.

Most of the methods of presentation used are the common property of many mathematicians, but I wish to acknowledge that the way of introducing compactness is influenced by Hewitt (1960). It is also a pleasure to acknowledge the influence of many teachers, colleagues, and ex-students on this book, and to thank Peter Strain of the Open University for helpful comments and the staff of the Clarendon Press for their encouragement during the writing.

Oxford, 1974 W.A.S.

Preface to reprinted edition

I am grateful to all who have pointed out errors in the first printing (even to those who pointed out that the proof of Corollary 1.1.7 purported to establish the existence of a *positive* rational number between any two real numbers). In particular, it is a pleasure to thank Roy Dyckhoff, Ioan James, and Richard Woolfson for valuable comments and corrections.

Oxford, 1981 W.A.S.

Contents

1 Introduction

In this book we are going to generalize theorems about convergence and continuity which are probably familiar to the reader in the case of sequences of real numbers and real-valued functions of one real variable. The kind of result we shall be trying to generalize is the following: *if a real-valued function f is defined and continuous on the closed interval $[a, b]$ in the real line, then f is bounded on $[a, b]$, i.e. there exists a real number K such that $|f(x)| \leqslant K$ for all x in $[a, b]$.* Several such theorems about real-valued functions of a real variable are true and useful in a more general framework, after suitable minor changes of wording. For example, if we suppose that a real-valued function f of *two* real variables is defined and continuous on a rectangle $[a, b] \times [c, d]$, then f is bounded on this rectangle. Once we have seen that the result generalizes from one to two real variables, it is natural to suspect that it is true for any finite number of real variables, and then to go a step further by asking: how general a situation can the theorem be formulated for, and how generally is it true? These questions lead us first to metric spaces and eventually to topological spaces.

Before going on to study such questions, it is fair to ask: what is the point of generalization? One answer is that it saves time, or at least avoids tedious repetition. If we can show by a single proof that a certain result holds for functions of n real variables, where n is any positive integer, this is better than proving it separately for one real variable, two real variables, three real variables, etc. In the same vein, generalization often gives a unified mental grasp of several results which otherwise might just seem vaguely similar, and in addition to the satisfaction involved, this more efficient organization of material helps some people's understanding. Another gain is that generalization often illuminates the proof of a theorem, because to see how generally a given result can be proved, one has to notice exactly which properties or hypotheses are used at each stage in the proof.

Against this, we should be aware of some dangers in generalization. Most mathematicians would agree that it can be carried to an excessive extent. Just when this stage is reached is a matter of controversy, but the potential reader is warned that some mathematicians would say 'Enough,

no more (at least as far as analysis is concerned)' when we get into metric spaces. Also, there is an initial barrier of unfamiliarity to be overcome in moving to a more general framework, with its new language; the extent to which the pay-off is worthwhile is likely to vary from one student to another.

Our successive generalizations lead to the subject called topology. Applications of topology range from analysis, geometry, and number theory to mathematical physics and computer science. Topology is a language for many mathematical topics, just as mathematics is a language for many sciences. But it also has attractive results of its own. We have mentioned that some of these generalize theorems the reader has already met for real-valued functions of a real variable. Moreover, topology has a geometric aspect which is familiar in popular expositions as 'rubber-sheet geometry', with pictures of doughnuts, Möbius bands, Klein bottles, and the like; we touch on this in the chapter on quotients, trying to indicate how such topics are part of the same story as the more analytic aspects. From the point of view of analysis, topology is the study of continuity, while from the point of view of geometry, it is the study of those properties of geometric objects which are preserved when the objects are stretched, compressed, bent, and otherwise mistreated—everything is legitimate except tearing apart and sticking together. This is what gives rise to the old joke that a topologist is a person who cannot tell the difference between a coffee cup and a doughnut the point being that each of these is a solid object with just one hole through it.

As a consequence of introducing abstractions gradually, the theorem density in this book is low. The title of theorem is reserved for substantial results, which have significance in a broad range of mathematics.

Some exercises are marked $\star$ or even $\star\star$ and some passages are enclosed between $\bigstar$ signs to denote that they are tentatively thought to be more challenging than the rest. A few paragraphs are enclosed between $\blacktriangleright$ and $\blacktriangleleft$ signs to denote that they require some knowledge of abstract algebra.

We shall try to illustrate the exposition with suitable diagrams; in addition readers are urged to draw their own diagrams wherever possible.

A word about the exercises: there are lots. Rather than being daunted, try a sample at a first reading, some more on revision, and so on. Hints are given with some of the exercises, and there are further hints on the web site. When you have done most of the exercises you will have an excellent understanding of the subject.

A previous course in real analysis is a prerequisite for reading this book. This means an introduction (including rigorous proofs) to continuity,

differential and preferably also integral calculus for real-valued functions of one real variable, and convergence of real number sequences. This material is included, for example, in Hart (2001) or, in a slightly more sophisticated but very complete way, in Spivak (2006) (names followed by dates in parentheses refer to the bibliography at the end of the book). The experience of abstraction gained from a previous course, in say, linear algebra, would help the reader in a general way to follow the abstraction of metric and topological spaces. However, the student is likely to be the best judge of whether he/she is ready, or wants, to read this book.

2 Notation and terminology

We use the logical symbols $\Rightarrow$ and $\Leftrightarrow$ meaning *implies* and *if and only if*. We also use iff to mean 'if and only if'; although not pretty, it is short and we use it frequently. Most introductions to algebra and analysis survey many parts of the language of sets and maps, and for these we just list notation.

If an object a belongs to a set A we write $a \in A$, or occasionally $A \ni a$, and if not we write $a \notin A$. If A is a subset of B (perhaps equal to B) we write $A \subseteq B$, or occasionally $B \supseteq A$. The subset of elements of A possessing some property P is written $\{a \in A : P(a)\}$. A finite set is sometimes specified by listing its elements, say $\{a_1, a_2, \ldots, a_n\}$. A set containing just one element is called a *singleton* set. Intersection and union of sets are denoted by $\cap$, $\cup$, or $\bigcap$, $\bigcup$. The empty set is written $\emptyset$. If $A \cap B = \emptyset$ we say that A and B are *disjoint*. Given two sets A and B, the set of elements which are in B but not in A is written $B \setminus A$. Thus in particular if $A \subseteq B$ then $B \setminus A$ is the complement of A in B. If S is a set and for each i in some set I we are given a subset A_i of S, then we denote by $\bigcup_{i \in I} A_i$, $\bigcap_{i \in I} A_i$ (or just $\bigcup_I A_i$, $\bigcap_I A_i$) the union and intersection of the A_i over all $i \in I$; for example, in the case of union what this means is

$$s \in \bigcup_{i \in I} A_i, \Leftrightarrow \text{ there exists } i \in I \text{ such that } s \in A_i.$$

In this situation I is called an *indexing* set. We use De Morgan's laws, which with the above notation assert

$$S \setminus \bigcup_I A_i = \bigcap_I (S \setminus A_i), \qquad S \setminus \bigcap_I A_i = \bigcup_I (S \setminus A_i).$$

In particular, if the indexing set is the positive integers $\mathbb{N}$ we usually write

$$\bigcup_{i=1}^{\infty} A_i, \quad \bigcap_{i=1}^{\infty} A_i \quad \text{for} \quad \bigcup_{i \in \mathbb{N}} A_i, \quad \bigcap_{i \in \mathbb{N}} A_i.$$

The Cartesian product $A \times B$ of sets A, B is the set of all ordered pairs (a, b) where $a \in A$, $b \in B$. This generalizes easily to the product of any

finite number of sets; in particular we use A^n to denote the set of ordered n-tuples of elements from A.

A map or function f (we use the terms interchangably) between sets X, Y is written $f : X \to Y$. We call X the *domain* of f, and we avoid calling Y anything. We think of f as assigning to each x in X an element $f(x)$ in Y, although logically it is preferable to define a map as a pair of sets X, Y together with a certain type of subset of $X \times Y$ (intuitively the graph of f). Persisting with our way of thinking about f, we define *the graph of* f to be the subset $G_f = \{(x, y) \in X \times Y : f(x) = y\}$ of $X \times Y$.

We call $f : X \to Y$ *injective* if $f(x) = f(x') \Rightarrow x = x'$ (we prefer this to 'one–one' since the latter is a little ambiguous). We should therefore call $f : X \to Y$ surjective if for every $y \in Y$ there is an $x \in X$ with $f(x) = y$, but we usually call such an f *onto*. If $f : X \to Y$ is both injective and onto we call it *bijective* or *a one–one correspondence*.

If $f : X \to Y$ is a map and $A \subseteq X$ then the *restriction of* f *to* A, written $f|A$, is the map $f|A : A \to Y$ defined by $(f|A)(a) = f(a)$ for every $a \in A$. In traditional calculus the function $f|A$ would not be distinguished from f itself, but when we are being fussy about the precise domains of our functions it is important to make the distinction: f has domain X while $f|A$ has domain A.

If $f : X \to Y$ and $g : Y \to Z$ are maps then their *composition* $g \circ f$ is the map $g \circ f : X \to Z$ defined by $(g \circ f)(x) = g(f(x))$ for each $x \in X$. This is the abstract version of 'function of a function' that features, for example, in the chain rule in calculus.

There are some more concepts relating to sets and functions which we shall focus on in the next chapter.

We shall occasionally assume that the terms *equivalence relation* and *countable set* are understood.

We use $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ to denote the sets of positive integers, integers, rational numbers, real numbers, and complex numbers, respectively. We often refer to $\mathbb{R}$ as the *real line* and we call the following subsets of $\mathbb{R}$ *intervals*:

$$\begin{aligned}
&\text{(i)} && [a, b] = \{x \in \mathbb{R} : a \leqslant x \leqslant b\},\\
&\text{(ii)} && (a, b) = \{x \in \mathbb{R} : a < x < b\},\\
&\text{(iii)} && (a, b] = \{x \in \mathbb{R} : a < x \leqslant b\},\\
&\text{(iv)} && [a, b) = \{x \in \mathbb{R} : a \leqslant x < b\},\\
&\text{(v)} && (-\infty, b] = \{x \in \mathbb{R} : x \leqslant b\},\\
&\text{(vi)} && (-\infty, b) = \{x \in \mathbb{R} : x < b\},
\end{aligned}$$

$$\begin{aligned} &\text{(vii)} && [a, \infty) = \{x \in \mathbb{R} : x \geqslant a\}, \\ &\text{(viii)} && (a, \infty) = \{x \in \mathbb{R} : x > a\}, \\ &\text{(ix)} && (-\infty, \infty) = \mathbb{R}. \end{aligned}$$

This is our definition of *interval* a subset of $\mathbb{R}$ is an interval iff it is on the above list. The intervals in (i), (v), (vii) (and (ix)) are called *closed* intervals; those in (ii), (vi), (viii) (and (ix)) are called *open* intervals; and (iii), (iv) are called *half-open* intervals. When we refer to an interval of types (i)–(iv), it is always to be understood that $b > a$, except for type (i), when we also allow $a = b$. We shall try to avoid the occasional risk of confusing an interval (a, b) in $\mathbb{R}$ with a point (a, b) in $\mathbb{R}^2$ by stating which of these is meant when there might be any doubt.

The reader has probably already had practice working with sets; here as revision exercises are a few facts which appear later in the book. The last two exercises, involving equivalence relations, are relevant to the chapter on quotient spaces (and only there). They look more complicated than they really are.

Exercise 2.1 Suppose that C, D are subsets of a set X. Prove that

$$(X \setminus C) \cap D = D \setminus C.$$

Exercise 2.2 Suppose that A, V are subsets of a set X. Prove that

$$A \setminus (V \cap A) = A \cap (X \setminus V).$$

Exercise 2.3 Suppose that V, X, Y are sets with $V \subseteq X \subseteq Y$ and suppose that U is a subset of Y such that $X \setminus V = X \cap U$. Prove that

$$V = X \cap (Y \setminus U).$$

Exercise 2.4 Suppose that U, V are subsets of sets X, Y, respectively Prove that

$$U \times V = (X \times V) \cap (U \times Y).$$

Exercise 2.5 Suppose that U_1, U_2 are subsets of a set X and that V_1, V_2 are subsets of a set Y. Prove that

$$(U_1 \times V_1) \cap (U_2 \times V_2) = (U_1 \cap U_2) \times (V_1 \cap V_2).$$

Exercise 2.6 Suppose that for some set X and some indexing sets I, J we have $U = \bigcup_{i \in I} B_{i1}$ and $V = \bigcup_{j \in J} B_{j2}$ where each B_{i1}, B_{j2} is a subset of X. Prove that

$$U \cap V = \bigcup_{(i,j) \in I \times J} B_{i1} \cap B_{j2}.$$

Exercise 2.7 (a) Let $\sim$ be an equivalence relation on a set X. Show that the corresponding equivalence classes partition X into a union of pairwise disjoint non-empty subsets $\{A_i : i \in I\}$ for some indexing set I. (This means that for all $i, j \in I$, we have $A_i \subseteq X$, $A_i \neq \emptyset$, $A_i \cap A_j = \emptyset$ for $i \neq j$, and $\bigcup_{i \in I} A_i = X$).

(b) Conversely show that a partition of X into pairwise disjoint non-empty subsets, say $\mathcal{P} = \{A_i : i \in I\}$, determines an equivalence relation $\sim$ on X where $x_1 \sim x_2$ iff x_1 and x_2 belong to the same set A_i in $\mathcal{P}$.

Exercise 2.8 Continuing with the notation of Exercise 2.7, let the partition determined by an equivalence relation $\sim$ on X be denoted by $\mathcal{P}(\sim)$ and the equivalence relation determined by a partition $\mathcal{P}$ be denoted by $\sim(\mathcal{P})$. Show that $\sim(\mathcal{P}(\sim)) = \sim$ and $\mathcal{P}(\sim(\mathcal{P})) = \mathcal{P}$. This shows that there is a one one correspondence between equivalence relations on X and partitions of X.

3 More on sets and functions

In the previous chapter we assumed familiarity with a certain amount of notation and terminology about sets and functions; but some readers may not yet be as much at ease with the concepts in the present chapter. In topology the idea of the *inverse image* of a set under a map is much used, so it is good to be familiar with it. If you are at ease with Definitions 3.1 and 3.2 below, then you could safely skip the rest of this chapter. (If in doubt, skip it now but come back to it later if necessary.)

Direct and inverse images

Let $f : X \to Y$ be any map, and let A, C be subsets of X, Y respectively.

Definition 3.1 *The* (direct *or* forwards) image $f(A)$ *of* A *under* f *is the subset of* Y *given by* $\{y \in Y : y = f(a) \text{ for some } a \in A\}$.

Definition 3.2 *The* inverse image $f^{-1}(C)$ *of* C *under* f *is the subset of* X *given by* $\{x \in X : f(x) \in C\}$.

We note immediately that in order to make sense Definition 3.2 does *not* require the existence of an 'inverse function' f^{-1}. *Pre-image* is possibly a safer name, but *inverse image* is more common so we shall stick to it. For the same reason, to avoid confusion with inverse functions, at least one textbook has very reasonably tried to popularize the notation $f^{\dashv}(C)$ in place of $f^{-1}(C)$, but this has not caught on, so we shall grasp the nettle and use $f^{-1}(C)$.

A particularly confusing case is $f^{-1}(y)$ for $y \in Y$. The confusion is enhanced by the notation: $f^{-1}(y)$ should really be written $f^{-1}(\{y\})$. It is the special case of $f^{-1}(C)$ when C is the singleton set $\{y\}$. We shall see examples below in which $f^{-1}(y)$ contains more than one element. We follow common usage by writing $f^{-1}(y)$ for $f^{-1}(\{y\})$ except in the next example.

Example 3.3 Let $X = \{x, y, z\}$, $Y = \{1, 2, 3\}$ and define $f : X \to Y$ by $f(x) = 1$, $f(y) = 2$, $f(z) = 1$. Then we have $f(\{x, y\}) = \{1, 2\}$, $f(\{x, z\}) = \{1\}$, $f^{-1}(\{1\}) = \{x, z\}$, and $f^{-1}(\{2, 3\}) = \{y\}$.

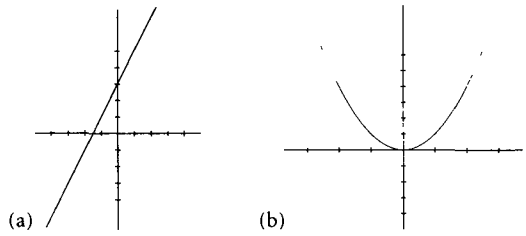

Figure 3.1. (a) Graph of f and (b) graph of g

As mentioned, we henceforth write $f^{-1}(\{1\})$ as $f^{-1}(1)$. Note $f^{-1}(1)$ here is *not* a singleton set.

Example 3.4 Let $X = Y = \mathbb{R}$ and define $f : X \to Y$ by $f(x) = 2x + 3$. The graph of this function is a straight line (see Figure 3.1(a)):

Then for example,

$$f([0, 1]) = [3, 5],\ f((1, \infty)) = (5, \infty),\ \ f^{-1}([0, 1]) = [-3/2, -1].$$

Example 3.5 Again let $X = Y = \mathbb{R}$. Define g by $g(x) = x^2$. The graph of this function has the familiar parabolic shape as in Figure 3.1(b). Then for example,

$$g([0, 1]) = [0, 1],\ \ g([1, 2]) = [1, 4],\ \ g(\{-1, 1\}) = \{1\},$$
$$g^{-1}([0, 1]) = [-1, 1],\ g^{-1}([1, 2]) = [-\sqrt{2}, -1] \cup [1, \sqrt{2}],\ g^{-1}([0, \infty)) = \mathbb{R}.$$

The special case of direct image and inverse image of the empty set are worth noting: for any map $f : X \to Y$ we have $f(\emptyset) = \emptyset$ and $f^{-1}(\emptyset) = \emptyset$: for example, $f^{-1}(\emptyset)$ consists of all elements of X which are mapped by f into the empty set, and there are no such elements so $f^{-1}(\emptyset) = \emptyset$.

We now come to some important formulae involving direct and inverse images. We state those about unions and intersections first in the case of just two subsets.

Proposition 3.6 *Suppose that $f : X \to Y$ is a map, that A, B are subsets of X and that C, D are subsets of Y. Then:*

$$f(A \cup B) = f(A) \cup f(B),\ \ f(A \cap B) \subseteq f(A) \cap f(B),$$
$$f^{-1}(C \cup D) = f^{-1}(C) \cup f^{-1}(D),\ \ f^{-1}(C \cap D) = f^{-1}(C) \cap f^{-1}(D).$$

Equality does not necessarily hold in the second formula, as we shall see shortly. There is a more general form of Proposition 3.6.

Proposition 3.7 *Suppose that $f : X \to Y$ is a map, and that for each i in some indexing set I we are given a subset A_i of X and a subset C_i of Y. Then*

$$f\left(\bigcup_{i\in I} A_i\right) = \bigcup_{i\in I} f(A_i), \qquad f\left(\bigcap_{i\in I} A_i\right) \subseteq \bigcap_{i\in I} f(A_i),$$

$$f^{-1}\left(\bigcup_{i\in I} C_i\right) = \bigcup_{i\in I} f^{-1}(C_i), \qquad f^{-1}\left(\bigcap_{i\in I} C_i\right) = \bigcap_{i\in I} f^{-1}(C_i).$$

As a sample of the proof we show that

$$f^{-1}\left(\bigcap_{i\in I} C_i\right) = \bigcap_{i\in I} f^{-1}(C_i).$$

(Proofs of the other parts of Proposition 3.7 are on the web site.) First let $x \in f^{-1}\left(\bigcap_{i\in I} C_i\right)$. Then $f(x) \in \bigcap_{i\in I} C_i$, so $f(x) \in C_i$ for every $i \in I$. This tells us that $x \in f^{-1}(C_i)$ for every $i \in I$, so $x \in \bigcap_{i\in I} f^{-1}(C_i)$. Hence,

$$f^{-1}\left(\bigcap_{i\in I} C_i\right) \subseteq \bigcap_{i\in I} f^{-1}(C_i).$$

The reverse inclusion is proved by running the argument backwards. Explicitly, if $x \in \bigcap_{i\in I} f^{-1}(C_i)$ then for every $i \in I$ we have $x \in f^{-1}(C_i)$, so $f(x) \in C_i$. This tells us that $f(x) \in \bigcap_{i\in I} C_i$, so $x \in f^{-1}\left(\bigcap_{i\in I} C_i\right)$ as required.

Next we give results about complements, again preceded by a special case.

Proposition 3.8 *Suppose that $f : X \to Y$ is a map and $B \subseteq X, D \subseteq Y$. Then*

$$f(X \setminus B) \supseteq f(X) \setminus f(B), \quad f^{-1}(Y \setminus D) = X \setminus f^{-1}(D).$$

This follows by taking $A = X$, $C = Y$ in the next proposition (for the second part of Proposition 3.8 we use also $f^{-1}(Y) = X$).

Proposition 3.9 *With the notation of Proposition 3.6,*

$$f(A \setminus B) \supseteq f(A) \setminus f(B) \quad \text{and} \quad f^{-1}(C \setminus D) = f^{-1}(C) \setminus f^{-1}(D).$$

The proof is on the web site.

We now explore Propositions 3.6 and 3.8 further, in order to gain familiarity. Here are two examples in which $f(A \cap B) = f(A) \cap f(B)$ fails and one in which $f(A \setminus B) = f(A) \setminus f(B)$ fails.

Example 3.10 Let $X = \{a, b\}$, $Y = \{1, 2\}$ and $f(a) = 1$, $f(b) = 1$. Put $A = \{a\}$, $B = \{b\}$. Then $A \cap B = \emptyset$, so $f(A \cap B) = \emptyset$. But on the other hand $f(A) \cap f(B) = \{1\} \neq \emptyset$.

Example 3.11 Let $X = Y = \mathbb{R}$, define $g(x) = x^2$, and let $A = [0, 1)$, $B = (-1, 0]$ so that $A \cap B = \{0\}$. Then $g(A \cap B) = \{0\}$ but on the other hand $g(A) \cap g(B) = [0, 1)$.

Example 3.12 Let $X = \{x, y, z\}$, $Y = \{1, 2, 3\}$ and as in Example 3.3 let $f(x) = 1 = f(z)$, $f(y) = 2$. Put $B = \{z\}$. Then $f(X \setminus B) = \{1, 2\}$, but on the other hand $f(X) \setminus f(B) = \{2\}$

The next result is useful later.

Proposition 3.13 *Suppose that $f : X \to Y$ is a map, $B \subseteq Y$ and for some indexing set I there is a family $\{A_i : i \in I\}$ of subsets of X with $X = \bigcup_I A_i$. Then*

$$f^{-1}(B) = \bigcup_I (f|A_i)^{-1}(B).$$

Proof First suppose $x \in f^{-1}(B)$. Since $X = \bigcup_I A_i$ we have $x \in A_{i_0}$ for some $i_0 \in I$. Then $(f|A_{i_0})(x) = f(x) \in B$, so $x \in (f|A_{i_0})^{-1}(B)$, which is contained in $\bigcup_I (f|A_i)^{-1}(B)$.

Conversely suppose that $x \in \bigcup_I (f|A_i)^{-1}(B)$. Then $x \in (f|A_{i_0})^{-1}(B)$ for some $i_0 \in I$. This says $(f|A_{i_0})(x) \in B$. But $(f|A_{i_0})(x) = f(x)$, so $f(x) \in B$ which gives $x \in f^{-1}(B)$. □

★ We occasionally want to look at sets such as $f^{-1}(f(A))$ or $f(f^{-1}(C))$; we look at a few basic facts about these, and explore them further in the exercises.

Proposition 3.14 *Let X, Y be sets and $f : X \to Y$ a map. For any subset $C \subseteq Y$ we have $f(f^{-1}(C)) = C \cap f(X)$. In particular, $f(f^{-1}(C)) = C$ if f is onto. For any subset $A \subseteq X$ we have $A \subseteq f^{-1}(f(A))$.*

Proof First let $y \in f(f^{-1}(C))$. Then $y = f(x)$ for some $x \in f^{-1}(C)$. But for such an x we have $f(x) \in C$, so $y \in C$. But also $y = f(x)$ so $y \in f(X)$. Hence $y \in C \cap f(X)$ and we have proved $f(f^{-1}(C)) \subseteq C \cap f(X)$. Suppose conversely that $y \in C \cap f(X)$. Then $y \in C$, and also $y = f(x)$ for some $x \in X$. Now for this x we have $f(x) = y \in C$, so $x \in f^{-1}(C)$. So $y = f(x) \in f(f^{-1}(C))$ as required, and we have proved the reverse inclusion $C \cap f(X) \subseteq f(f^{-1}(C))$. Thus $f(f^{-1}(C)) = C \cap f(X)$. When f is onto, $f(X) = Y$ so $f(f^{-1}(C)) = C$.

Secondly, for any $a \in A$ we have $f(a) \in f(A)$ so $a \in f^{-1}(f(A))$ as required. □

It is easy to find examples where the inclusion in the last part is strict.

Example 3.15 Following Example 3.10 let $X = \{a, b\}$, $Y = \{1, 2\}$, and $f(a) = 1 = f(b)$, $A = \{a\}$. Then $f^{-1}(f(A)) = f^{-1}(1) = \{a, b\} \neq A$.

Example 3.16 Let $X = Y = \mathbb{R}$ and let $g(x) = x^2$. Put $A = [0, 1]$. Then $g^{-1}(g(A)) = g^{-1}([0, 1]) = [-1, 1] \neq A$. ★

Inverse functions

We have emphasized that in order for the inverse image $f^{-1}(C)$ to be defined, there need not exist any inverse function f^{-1}. We now look at the case when such an inverse does exist.

Definition 3.17 *A map $f : X \to Y$ is said to be* invertible *if there exists a map $g : Y \to X$ such that the composition $g \circ f$ is the identity map of X and the composition $f \circ g$ is the identity map of Y.*

We immediately get a criterion on f for it to be invertible:

Proposition 3.18 *A map $f : X \to Y$ is invertible if and only if it is bijective.*

Proof Suppose first that f is invertible and let g be as in Definition 3.17. Then

$$f(x) = f(x') \Rightarrow g(f(x)) = g(f(x')) \Rightarrow x = x'$$

so f is injective. Also, given any $y \in Y$ we have $y = f(g(y))$ so $y \in f(X)$, which says that f is onto. Hence f is bijective.

Secondly suppose that f is bijective. We may define $g : Y \to X$ as follows: for any $y \in Y$ we know f is onto, so $y = f(x)$ for some $x \in X$. Moreover this x is unique for a given y since f is injective. Put $g(y) = x$, and we can see that f and g satisfy Definition 3.17, so f is invertible as required. □

The last part of the above proof also proves

Proposition 3.19 *When f is invertible, there is a* unique *g satisfying Definition 3.17. This unique g is called* the inverse of f, *written f^{-1}.*

For given $y \in Y$, in order to satisfy Definition 3.17 we have to choose $g(y)$ to be the unique $x \in X$ such that $f(x) = y$.

The final result in this chapter is slightly tricky, but it is very useful for one important theorem later (Theorem 13.26).

Proposition 3.20 *Suppose that $f : X \to Y$ is a one–one correspondence of sets X and Y and that $V \subseteq X$. Then the inverse image of V under the inverse map $f^{-1} : Y \to X$ equals the image set $f(V)$.*

Proof Let us write $g : Y \to X$ for the inverse function f^{-1} of $f : X \to Y$. We want to show for any $V \subseteq X$ that $g^{-1}(V) = f(V)$.

First suppose y is in $f(V)$. Then $y = f(x)$ for some $x \in V$, and this x is unique since f is injective. By definition of inverse function $x = g(y)$. But since $x \in V$ this gives $y \in g^{-1}(V)$. We have now proved $f(V) \subseteq g^{-1}(V)$.

Secondly suppose $y \in g^{-1}(V)$. Then $g(y) \in V$. So $f(g(y)) \in f(V)$. But g is the inverse function to f, so $f(g(y)) = y$, and we have $y \in f(V)$. This shows that $g^{-1}(V) \subseteq f(V)$. So we have proved $g^{-1}(V) = f(V)$ as required.

We may write the conclusion in the following rather mind-boggling way: $(f^{-1})^{-1}(V) = f(V)$. The inner superscript -1 indicates the function f^{-1} is inverse to f, and the outer one indicates the inverse image of the set V under that inverse function. □

Although some textbooks write f^{-1} only when f is invertible, others take the more relaxed view that if $f : X \to Y$ is injective, then it defines a bijective function $f_1 : X \to f(X)$, and they write $f^{-1} : f(X) \to X$ for the inverse of f_1 in the sense of Definition 3.17 and Proposition 3.19 above. This is a useful alternative, although we shall stick to the narrower interpretation.

Of the exercises, 3.5, 3.6, and 3.9 involve the starred section above.

Exercise 3.1 Let $f : X \to Y$ be a map and suppose that $A \subseteq B \subseteq X$ and that $C \subseteq D \subseteq Y$. Prove that $f(A) \subseteq f(B) \subseteq Y$ and that $f^{-1}(C) \subseteq f^{-1}(D) \subseteq X$

Exercise 3.2 Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = \sin x$. Describe the sets:

$$f([0, \pi/2]).\ f([0.\infty)),\ f^{-1}([0,\ 1]),\ f^{-1}([0.\ 1/2]).\ f^{-1}([-1,\ 1]).$$

Exercise 3.3 Suppose that $f : X \to Y$ and $g : Y \to Z$ are maps and $U \subseteq Z$. Prove that $(g \circ f)^{-1}(U) = f^{-1}(g^{-1}(U))$

Exercise 3.4 Let $f\ .\ \mathbb{R} \to \mathbb{R}^2$ be defined by $f(x) = (x.\ 2x)$ Describe the sets:

$$f([0.\ 1]).\ f^{-1}([0,\ 1] \times [0,\ 1]).\ f^{\ -1}(D) \text{ where } D = \{(x,\ y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant 1\}$$

Exercise 3.5 Show that a map $f : X \to Y$ is onto iff $f(f^{-1}(C)) = C$ for all subsets $C \subseteq Y$.

Exercise 3.6 Show that a map $f : X \to Y$ is injective iff $A = f^{-1}(f(A))$ for all subsets $A \subseteq X$.

Exercise 3.7 Let $f \cdot X \to Y$ be a map. For each of the following determine whether it is true in general or whether it is sometimes false (Give a proof or a counterexample for each.)

(i) If $y.\ y' \in Y$ with $y \neq y'$ then $f^{-1}(y) \neq f^{-1}(y')$.

(ii) If $y,\ y' \in Y$ with $y \neq y'$ and f is onto then $f^{-1}(y) \neq f^{-1}(y')$.

Exercise 3.8 Let $f : X \to Y$ be a map and let $A,\ B$ be subsets of X. Prove that $f(A \setminus B) = f(A) \setminus f(B)$ if and only if $f(A \setminus B) \cap f(B) = \emptyset$ Deduce that if f is injective then $f(A \setminus B) = f(A) \setminus f(B)$.

Exercise 3.9 Let $f : X \to Y$ be a map and $A \subseteq X,\ C \subseteq Y$. Prove that

(a) $f(A) \cap C = f(A \cap f^{-1}(C))$.

(b) if also $B \subseteq X$ and $f^{-1}(f(B)) = B$ then $f(A) \cap f(B) = f(A \cap B)$.

Exercise 3.10 Suppose that $f : X \to Y$ is a map from a set X onto a set Y. Show that the family of subsets $\{f^{-1}(y) : y \in Y\}$ forms a partition of X in the sense of Exercise 2.7.

4 Review of some real analysis

The point of this chapter is to review a few basic ideas in real analysis which will be generalized in later chapters. It is not intended to be an introduction to these concepts for those who have never seen them before.

Real numbers

Two popular ways of thinking about the real number system are:

(1) geometrically, as corresponding to all the points on a straight line;

(2) in terms of decimal expansions, where if a number is irrational we think of longer and longer decimal expansions approximating it more and more closely.

Neither of these intuitive ideas is precise enough for our purposes, although each leads to a way of constructing the real numbers from the rational numbers. The second of these ways is described on the web site.

One approach to real numbers is axiomatic. This means we write down a list of properties and define the real numbers to be any system satisfying these properties. The properties are called *axioms* when they are used in this way. Another approach is constructive: we construct the real numbers from the rationals. The rational numbers may in turn be constructed from the integers, and so on—we can follow the trail backwards through the positive integers and back to set theory. (One has to begin with axioms at some stage, however.) In either approach the set of real numbers has certain properties; depending on the approach we have in mind, we call these properties either axioms or propositions. We shall assume that the construction of $\mathbb{R}$ has already been carried out for us, and we are interested in its properties.

Many introductions to analysis contain a list of properties of real numbers (see, for example, Hart (2001) or Spivak (2006)). A large number of these may be summed up technically by saying that the real numbers form an ordered field. Roughly this means that addition, subtraction, multiplication, and division of real numbers all work in the way we expect them to, and that the same is true of the way in which inequalities $x < y$ work and interact with addition and multiplication. We shall not review these properties, but concentrate on the so-called completeness property. The reasons for this strange behaviour are, first, that this is the property

which distinguishes the real numbers from the rational numbers (and in a sense analysis and topology from algebra) and secondly that our intuition is unlikely to let us down on properties deducible from those of an ordered field, whereas arguments using completeness tend to be more subtle.

To state the completeness property we need some terminology. Let S be a non-empty set of real numbers. An *upper bound for* S is a number x such that $y \leqslant x$ for all y in S. If an upper bound for S exists we say that S is bounded above. Lower bounds are defined similarly.

Example 4.1 (a) The set $\mathbb{R}$ of all real numbers has no upper or lower bound.

(b) The set $\mathbb{R}_-$ of all strictly negative real numbers has no lower bound, but for example 0 is an upper bound (as is any positive real number).

(c) The half-open interval $(0, 1]$ is bounded above and below.

If S has an upper bound u, then S has (infinitely) many upper bounds, since any $x \in \mathbb{R}$ satisfying $x \geqslant u$ is also an upper bound. This gives the next definition some point.

Definition 4.2 *Given a non-empty subset S of $\mathbb{R}$ which is bounded above, we call u a* least upper bound for S *if*

(a) u is an upper bound for S,

(b) $x \geqslant u$ for any upper bound x for S.

Example 4.3 In Example 4.1 (b), 0 is a least upper bound for $\mathbb{R}_-$. For 0 is an upper bound, and it is a *least* upper bound because any $x < 0$ is not an upper bound for $\mathbb{R}_-$ (since any such x satisfies $x/2 > x$ and $x/2 \in \mathbb{R}_-$). Examples 4.1 (c) and (b) show that a least upper bound of a set S may or may not be in S.

It follows from Definition 4.2 that least upper bounds are unique when they exist. For if u, u' are both least upper bounds for a set S, then since u' is an upper bound for S it follows that $u \leqslant u'$ by leastness of u ('leastness' means the property in Definition 4.2 (b)). Interchanging the roles of u and u' in this argument shows that also $u' \leqslant u$, so $u' = u$.

Greatest lower bounds are defined similarly to least upper bounds.

We can now state one form of the completeness property for $\mathbb{R}$.

Proposition 4.4 *Any non-empty subset of $\mathbb{R}$ which is bounded above has a least upper bound.*

Since our interest is in generalizing real analysis rather than studying its foundations, we offer no proof of Proposition 4.4. The completeness

property is quite subtle, and it is difficult to grasp its full significance until it has been used several times. It corresponds to the intuitive idea that there are no gaps in the real numbers, thought of as the points on a straight line; but the transition from the intuitive idea to the formal statement is not immediately obvious. For some sets of real numbers, such as Examples 4.1 (b) and (c), it is 'obvious' that a least upper bound exists (strictly speaking, this means that it follows from the properties of an ordered field). But this is not the case for all bounded non-empty sets of real numbers—for example, consider $S = \{x \in \mathbb{Q} : x^2 < 2\}$: the least upper bound turns out to be $\sqrt{2}$, and we need Proposition 4.4 to establish its existence indeed, the existence of $\sqrt{2}$ cannot follow from the ordered field properties alone, since $\mathbb{Q}$ is an ordered field, but there is no rational number whose square is 2 (see Exercise 4.5).

For any non-empty subset S of $\mathbb{R}$ which is bounded above we call its unique least upper bound $\sup S$ (sup is short for supremum). Other notation sometimes used is l.u.b. S.

Although the completeness property was stated in terms of sets bounded above, it is equivalent to the corresponding property for sets bounded below. The next proposition formally states half of this equivalence.

Proposition 4.5 *If a non-empty subset S of $\mathbb{R}$ is bounded below then it has a greatest lower bound.*

Proof Let $T = \{x \in \mathbb{R} : -x \in S\}$. The idea of the proof is simply that l is a lower bound for S if and only if $-l$ is an upper bound for T. The details are left as Exercise 4.7. □

Just as in the case of least upper bounds, a non-empty subset S of $\mathbb{R}$ which is bounded below has a unique greatest lower bound called $\inf S$ (short for infimum) or g.l.b. S.

The next proposition and its corollary are applications of the completeness property.

Proposition 4.6 *The set $\mathbb{N}$ of positive integers is not bounded above.*

Proof Suppose for a contradiction that $\mathbb{N}$ is bounded above. Then by the completeness property there is a real number $u = \sup \mathbb{N}$. For any $n \in \mathbb{N}$, $n + 1$ is also in $\mathbb{N}$, so $n + 1 \leqslant u$. But then $n \leqslant u - 1$. Hence $n \leqslant u - 1$ for any $n \in \mathbb{N}$, so $u - 1$ is an upper bound for $\mathbb{N}$, contradicting the leastness of u. This contradiction shows that $\mathbb{N}$ cannot be bounded above. □

Corollary 4.7 *Between any two distinct real numbers x and y there is a rational number.*

Proof Suppose first that $0 \leqslant x < y$. Since $y - x > 0$, by Proposition 4.6 there is an n in $\mathbb{N}$ such that $n > 1/(y-x)$ and hence $1/n < y - x$. Let $M = \{m \in \mathbb{N} : m/n > x\}$. By Proposition 4.6 $M \neq \emptyset$, otherwise nx would be an upper bound for $\mathbb{N}$. Hence, since $M \subseteq \mathbb{N}$, M contains a least integer m_0. So $m_0/n > x$ and $(m_0 - 1)/n \leqslant x$, from which $m_0/n \leqslant x + 1/n$. Hence $x < m_0/n \leqslant x + 1/n < x + (y - x) = y$, and m_0/n is a suitable rational number, between x and y. Now suppose that $x < 0$. If $y > 0$ then 0 is a rational number between x and y, while if $y \leqslant 0$ then the first case supplies a rational number r such that $-y < r < -x$, so $x < -r < y$ which says that the rational number $-r$ is between x and y. □

The above proofs of Proposition 4.6 and Corollary 4.7 assume several 'obvious' facts about $\mathbb{R}$ which we should really prove beforehand. For example, we deduced $n \leqslant u - 1$ from $n + 1 \leqslant u$, a consequence of the property often stated as follows: if a, b, $c \in \mathbb{R}$ and $a \leqslant b$ then $a + c \leqslant b + c$. Also, we assumed that any non-empty subset of $\mathbb{N}$ has a least element. We leave the reader to spot other such assumptions.

Remark 4.8 *Between any two distinct real numbers there is also an irrational number (see Exercise 4.8).*

We conclude this brief review of real numbers by recalling two useful inequalities, often called the triangle inequality and the reverse triangle inequality. There are proofs on the web site.

Proposition 4.9 $|x + y| \leqslant |x| + |y|$ *for any* x, y *in* $\mathbb{R}$.

Corollary 4.10 $|x - y| \geqslant \big||x| - |y|\big|$ *for any* x, y *in* $\mathbb{R}$.

Real sequences

Formally an infinite sequence of real numbers is a map $s : \mathbb{N} \to \mathbb{R}$. This definition is useful for discussing topics such as subsequences and rearrangements without being vague. In practice, however, given such a map s we denote $s(n)$ by s_n and think of the sequence in the traditional way as an infinite ordered string of numbers, using the notation (s_n) or $s_1, s_2, s_3, \ldots$ for the whole sequence.

It is important to distinguish between a sequence (s_n) and the set of its members $\{s_n : n \in \mathbb{N}\}$. The latter can easily be finite. For example if (s_n) is 1, 0, 1, 0, ... then its set of members is $\{0, 1\}$. Formally, this is a matter of distinguishing between a map $s : \mathbb{N} \to \mathbb{R}$ and its image set $s(\mathbb{N})$.

Sequences can arise, for example, in solving algebraic or differential equations. On the theoretical side, convergent sequences might be used to prove the existence of solutions to equations. On the practical side, s_n might be the answer at the nth stage in some method of successive approximations for finding a root of an equation. The only difference between theory and practice here is that in practice one is interested in how quickly the sequence gives a good approximation to the answer. Also, in applications we might be dealing with a sequence of vectors or of functions instead of real numbers.

We now review real number sequences, emphasizing those definitions and results whose analogues we shall later study for more general sequences.

Example 4.11 (a) $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \ldots,$ (*b*) $1, -\frac{1}{2}, \frac{1}{4}, -\frac{1}{8}, \ldots,$

(*c*) $\frac{1}{2}, 1, -\frac{1}{2}, -1, \frac{1}{4}, \frac{1}{2}, -\frac{1}{4}, -\frac{1}{2}, \ldots,$ (*d*) $1, 2, 3, \ldots,$ (*e*) $1, 0, 1, 0, \ldots,$

(f) $s_1 = 1,\ s_2 = 0,\ s_n = \frac{1}{2}(s_{n-2} + s_{n-1})$ for $n > 2,$

(g) s_n is the nth stage in some specified algorithm for approximating $\sqrt{2}$.

In examples (a)–(e), there is a simple formula for s_n in terms of n, which the reader will spot. This is convenient for illustrating the basic theory of sequences, but in practice a sequence might be generated by an iterative process, as in examples (f) and (g), or by the results of a probabilistic experiment repeated more and more often, or by some other means, and in such cases there may not be any simple formula for s_n in terms of n.

In Examples 4.11 (a), (b), (c) the sequence seems intuitively to be heading towards a definite number, whether steadily, or by alternately overshooting and undershooting the target, or irregularly, whereas in Examples 4.11 (d) (e) this is not the case. The mathematical term for 'heading towards' is 'converging', and the precise definition, as the reader probably knows, is as follows.

Definition 4.12 *The sequence* (s_n) converges to *(the real number)* l *if given (any real number)* $\varepsilon > 0$, *there exists (an integer)* N_ε *such that* $|s_n - l| < \varepsilon$ *for all* $n \geqslant N_\varepsilon$.

This is usually shortened by omitting the phrases in parentheses, and we often write just N in place of N_ε, although we need to remember that the value of N needed will usually vary with ε—intuitively, the smaller ε

is, the larger N will need to be. When Definition 4.12 holds, the number l is called the *limit* of the sequence. Other ways of writing '(s_n) converges to l' are '$s_n \to l$ as $n \to \infty$' and '$\lim_{n\to\infty} s_n = l$'. Here are two ways of thinking about the definition.

(1) (s_n) converges to l if, given any required degree of accuracy, then by going far enough along the sequence we can be sure that the terms beyond that stage all approximate l to within the required degree of accuracy.

(2) Let us take coordinate axes in the plane and mark the points with coordinates (n, s_n). Let us also draw a horizontal line L at height l. Then (s_n) converges to l if given any horizontal band of positive width centred on L, there exists a vertical line such that all marked points to the right of this vertical line lie within the prescribed horizontal band. Figure 4.1 is the kind of picture this suggests. The sequence promises to stay out of the shaded territory.

Two points are easy to get wrong when one is first trying to wield the formal definition. First, the order in which ε, N occur is crucial: given any $\varepsilon > 0$ *first*, there must *then* be an N_ε such that ... etc. Secondly, to prove convergence it is not enough to show that given $\varepsilon > 0$ there exists an N such that $|s_n - l| < \varepsilon$ for *some* $n \geqslant N$: this would be true of the sequence $1, 0, 1, 0, \ldots$, with $l = 0$, any $\varepsilon > 0$, and $N = 1$, yet the sequence does not converge.

The first deduction from the formal definition is an obvious part of the intuitive idea of convergence.

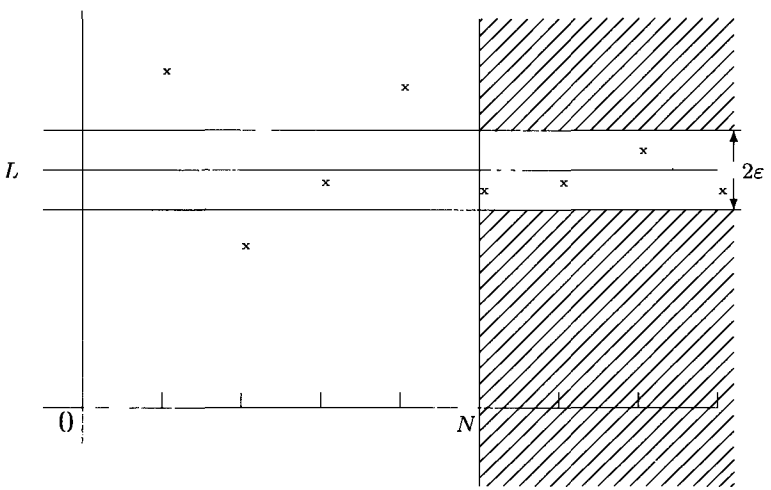

Figure 4.1. 'Graph' of a convergent sequence

Proposition 4.13 *A convergent sequence has a unique limit.*

Proof Suppose that (s_n) converges to l and also to l' where $l' \neq l$. Put $\varepsilon = \frac{1}{2}|l - l'|$. Since (s_n) converges to l, there is an integer N_1 such that $|s_n - l| < \varepsilon$ for all $n \geqslant N_1$. Similarly, since (s_n) converges to l', there is an integer N_2 such that $|s_n - l'| < \varepsilon$ for all $n \geqslant N_2$. Put $N = \max\{N_1, N_2\}$. Then, using the triangle inequality (Proposition 4.9),

$$|l - l'| = |l - s_N + s_N - l'| \leqslant |l - s_N| + |s_N - l'| < 2\varepsilon = |l - l'|.$$

This contradiction shows that $l' = l$. □

Before going further it is convenient to state explicitly a technical detail which is often used in convergence proofs.

Lemma 4.14 *Suppose there is a positive real number K such that given $\varepsilon > 0$ there exists N with $|s_n - l| < K\varepsilon$ for all $n \geqslant N$. Then (s_n) converges to l.*

Proof Let $\varepsilon > 0$. Then $\varepsilon/K > 0$, and if the stated condition holds, then there exists N such that $|s_n - l| < K(\varepsilon/K) = \varepsilon$ for all $n \geqslant N$, as required. In practice K is often an integer such as 2 or 3; we note that it needs to be independent of the choice of ε. □

In simple cases such as Example 4.11 (a) we can guess the limit and prove convergence directly. In general, however, it may be hard to guess the limit, and more importantly there may be no more convenient way to name a real number than as the limit of a given sequence. As an example consider:

$$s_n = 1 + \frac{1}{1!} + \frac{1}{2!} + \ldots + \frac{1}{n!}.$$

The reader may be able to think of a way to define the number e other than as the limit of the sequence (s_n), but it will also directly or indirectly involve taking the limit of this or some other sequence such as (t_n) where $t_n = (1 + 1/n)^n$.

We shall consider two theorems which provide ways of proving convergence without using a known value of the limit. As the above discussion indicates, both will depend heavily on the completeness property for $\mathbb{R}$.

Definition 4.15 *A sequence (s_n) is said to be* monotonic increasing (decreasing) *if $s_{n+1} \geqslant s_n$ ($s_{n+1} \leqslant s_n$) for all n in $\mathbb{N}$. It is* monotonic *if it has either of these properties.*

Theorem 4.16 *Every bounded monotonic sequence of real numbers converges.*

The proof is on the companion web site. As well as being useful on its own, Theorem 4.16 helps to prove the next convergence criterion. First we give a name to sequences in which the terms get closer and closer together as we get further along in the sequence.

Definition 4.17 *A sequence* (s_n) *is a* Cauchy *sequence if given* $\varepsilon > 0$ *there exists* N *such that if* $m, n \geqslant N$ *(i.e. if* $m \geqslant N$ *and* $n \geqslant N$*) then* $|s_m - s_n| < \varepsilon$.

Theorem 4.18 (Cauchy's convergence criterion) *A sequence* (s_n) *of real numbers converges if and only if it is a Cauchy sequence.*

Proof Suppose that (s_n) converges to l. Then given $\varepsilon > 0$, there exists N such that $|s_n - l| < \varepsilon$ for all $n \geqslant N$, so for $m, n \geqslant N$ the triangle inequality gives

$$|s_m - s_n| = |s_m - l + l - s_n| \leqslant |s_m - l| + |l - s_n| < 2\varepsilon.$$

Hence (s_n) is a Cauchy sequence (cf. Lemma 4.14).

Suppose conversely that (s_n) is a Cauchy sequence in $\mathbb{R}$. We show first that (s_n) is bounded. Take $\varepsilon = 1$, say, in the Cauchy condition. Thus there exists an N such that $m, n \geqslant N$ imply $|s_m - s_n| < 1$, so for any $m \geqslant N$ we have $|s_m - s_N| < 1$, and hence, using the triangle inequality,

$$|s_m| = |s_m - s_N + s_N| \leqslant |s_m - s_N| + |s_N| < 1 + |s_N|.$$

From this we get $|s_n| \leqslant \max\{|s_1|, |s_2|, \ldots |s_{N-1}|, 1 + |s_N|\}$ for all n, so (s_n) is bounded. (We could have used any fixed positive choice of ε in place of 1 in this part of the proof for example, 10^{10} or 10^{-10}.)

Next, in order to use Theorem 4.16, we manufacture a monotonic sequence out of (s_n) in the following subtle fashion. For each $m \in \mathbb{N}$ we let S_m be the set of members of the sequence from the mth stage onwards, $S_m = \{s_n : n \geqslant m\}$. Since the whole set of members $S = S_1$ of the sequence is bounded, so is S_m. Hence by the completeness property $\sup S_m$ exists. Let $t_m = \sup S_m$. Since $S_{m+1} \subseteq S_m$, we have $\sup S_{m+1} \leqslant \sup S_m$ (see Exercise 4.1). Thus the sequence (t_m) is monotonic decreasing. Also, $t_m \geqslant s_m$ by definition of t_m, and S is bounded below, so (t_m) is bounded below. So by Theorem 4.16, (t_m) converges, say to l.

Finally we prove, by a 3ε-argument, that (s_n) also converges to l. Given $\varepsilon > 0$ there exists N_1 such that $|s_m - s_n| < \varepsilon$ for $m, n \geqslant N_1$ and there

exists N_2 such that $|l-t_m| < \varepsilon$ for $m \geqslant N_2$. Put $N = \max\{N_1, N_2\}$. Since t_N is $\sup S_N$, we know that $t_N - \varepsilon$ is not an upper bound of S_N, so there exists $M \geqslant N$ such that $s_M > t_N - \varepsilon$; also, $s_M \leqslant t_N$ since $s_M \in S_N$ and t_N is an upper bound for S_N. Hence $|s_M - t_N| < \varepsilon$. Now for any $n \geqslant N$, using the triangle inequality twice,

$$|s_n - l| = |s_n - s_M + s_M - t_N + t_N - l| \leqslant |s_n - s_M| + |s_M - t_N| + |t_N - l| < 3\varepsilon.$$

Hence (s_n) converges to l (using Lemma 4.14). □

There is a further result about sequences which we record here for later reference: it is a version of the Bolzano–Weierstrass theorem.

Theorem 4.19 *Every bounded sequence of real numbers has at least one convergent subsequence.*

There is a proof on the web site.

Before leaving sequences we recall that their limits behave well under algebraic operations in the following sense.

Proposition 4.20 *Suppose that* (s_n), (t_n) *converge to* s, t. *Then*

(a) $(s_n + t_n)$ *converges to* $s + t$,

(b) $(s_n t_n)$ *converges to* st,

(c) $(1/t_n)$ *converges to* $1/t$ *provided* $t \neq 0$.

A few particular limits which we need are included in the exercises below.

Limits of functions

Limits of functions are used in the theoretical study of continuity, differentiability, and integration, and in practical estimates of the behaviour of particular functions.

Suppose first for simplicity that we have a function $f : \mathbb{R} \to \mathbb{R}$. (In general the domain could be smaller.) Let $a \in \mathbb{R}$.

Definition 4.21 *We say that* $f(x)$ tends to the limit l as x tends to a, *and write* $\lim_{x \to a} f(x) = l$, *if given (any real number)* $\varepsilon > 0$ *there exists (a real number)* $\delta > 0$ *such that* $|f(x) - l| < \varepsilon$ *for all real numbers* x *which satisfy* $0 < |x - a| < \delta$.

This is similar to the definition of convergence of a sequence (s_n), but instead of looking at s_n for large values of n, we look at $f(x)$ for x close to, **but not equal to**, a. Again the phrases in parentheses are usually omitted, and we note that the size of δ needed will in general depend

on ε. **The value $f(a)$ is irrelevant to the existence of $\lim_{x\to a} f(x)$, and the limit, if it exists, may or may not equal $f(a)$.** Exercise 4.12 is a good test of whether this important point has been fully absorbed.

Example 4.22 Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = x \quad \text{for } x \neq 0, \qquad f(0) = 1.$$

Then $\lim_{x\to 0} f(x) = 0$. For given $\varepsilon > 0$, put $\delta = \varepsilon$. If $0 < |x - 0| < \delta$, then $|f(x) - 0| = |x| < \varepsilon$, as required.

To emphasize further that $f(a)$ is irrelevant to the existence or value of $\lim_{x\to a} f(x)$, we note that Definition 4.21 makes sense even if $f(a)$ is not defined—it is enough to assume that f is defined on some subset $A \subseteq \mathbb{R}$, where A contains numbers arbitrarily close to (but not equal to) a. We shall not study this general case, but we note two especially useful ways of generalizing Definition 4.21. Suppose first that the domain A of f contains the open interval (a, d) for some $d > a$.

Definition 4.23 *The* right-hand limit $\lim_{x\to a+} f(x)$ *is equal to l if given $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(x) - l| < \varepsilon$ for all x in $(a, a + \delta)$.*

(Note that δ may be chosen small enough so that $(a, a + \delta) \subseteq (a, d)$, and therefore $f(x)$ is defined for all x in $(a, a + \delta)$.) Left-hand limits are defined similarly.

Next, here are two examples much used in illustrating theoretical points.

Example 4.24 Let $f, g : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ be given by

$$f(x) = x \sin 1/x, \qquad g(x) = \sin 1/x.$$

Then $\lim_{x\to 0} f(x) = 0$, while $\lim_{x\to 0} g(x)$ does not exist.

The proofs are left as Exercise 4.14.

Results about limits of functions may be proved by analogy with the proofs about sequences or we may deduce them from the latter using the following conversion lemma.

Lemma 4.25 *The following are equivalent:*

(i) $\lim_{x\to a} f(x) = l$,

(ii) if (x_n) is any sequence such that (x_n) converges to a but for all n we have $x_n \neq a$, then $(f(x_n))$ converges to l.

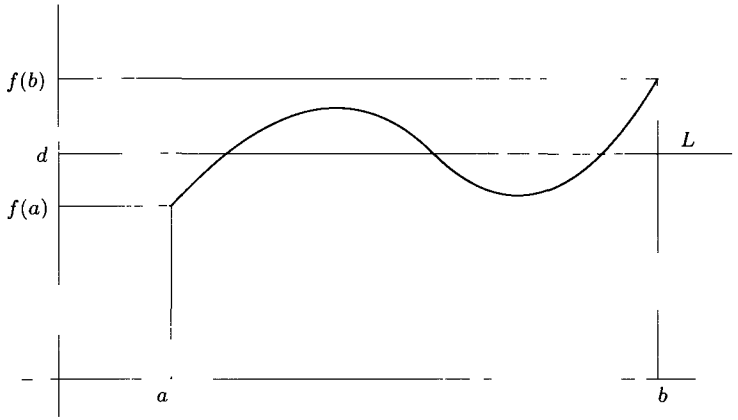

Figure 4.2. Intermediate value property

The proof is on the web site. One may also prove analogues of Theorem 4.18 and Proposition 4.20 for limits of functions, and for left- and right-hand limits.

Continuity

In this section we review the way in which a precise definition of continuity is derived from the intuitive notion. We first make a false start.

One statement containing something of the intuitive idea of continuity is that a function is continuous if its graph can be drawn without lifting pencil from paper. To formulate this more mathematically, let $f : \mathbb{R} \to \mathbb{R}$ be a function and let $(a, f(a))$, $(b, f(b))$ be two points on its graph (see Figure 4.2).

Let L be the horizontal line at some height d between $f(a)$ and $f(b)$. Then to satisfy our intuition about continuity, the graph of f has to cross the line L at least once on its way from $(a, f(a))$ to $(b, f(b))$. In other words, there exists at least one point c in $[a, b]$ such that $f(c) = d$. Formally, we make the following definition.

Definition 4.26 *A function $f : \mathbb{R} \to \mathbb{R}$ has the* intermediate value property (IVP) *if given any a, b, d in $\mathbb{R}$ with $a < b$ and d between $f(a)$ and $f(b)$, there exists at least one c satisfying $a \leqslant c \leqslant b$ and $f(c) = d$.*

This definition also applies when the domain $\mathbb{R}$ in Definition 4.26 is replaced by an interval in $\mathbb{R}$.

A tentative definition of continuity would be that f is continuous if it has the IVP. However, this fails to capture completely the intuitive idea of continuity, as the next example shows.

Example 4.27 Let f be given by

$$f(x) = \begin{cases} 0 & \text{for } x \leqslant 0, \\ \sin 1/x & \text{for } x > 0. \end{cases}$$

Part of the graph of f is shown in Figure 4.3.

Although we shall not prove it now, it is easy to believe by inspection that f does have the IVP. But f does not satisfy our intuitive requirements for a continuous function something is wrong near $x = 0$. On closer scrutiny, we realize that our intuition includes the requirement that for all values of x near 0, $f(x)$ should be reasonably close to $f(0)$, not oscillating with amplitude 1 as it does in this example. More precisely, the reason we are dissatisfied with f is that $\lim_{x \to 0} f(x)$ does not exist. Considerations such as these lead to the accepted definition.

Definition 4.28 *A function* $f : \mathbb{R} \to \mathbb{R}$ *is* continuous at a *if* $\lim_{x \to a} f(x)$ *exists and is* $f(a)$.

Using Definition 4.21 this translates into $\varepsilon - \delta$ form.

Definition 4.29 *A function* $f : \mathbb{R} \to \mathbb{R}$ *is* continuous at a *if given any* $\varepsilon > 0$, *there exists* $\delta > 0$ *such that* $|f(x) - f(a)| < \varepsilon$ *for any* x *such that* $|x - a| < \delta$.

As usual, the size of δ needed in general depends on ε, though we do not exhibit that in the notation.

A third way of expressing continuity of a function f is to say that $\lim_{x \to a-} f(x)$ and $\lim_{x \to a+} f(x)$ both exist and both equal $f(a)$. This has the advantage of identifying the ways in which continuity at a might fail: the left-hand limit, or the right-hand limit of f at a (or both of these) might fail to exist; or both left- and right-hand limits exist, but at least one of them fails to equal $f(a)$. (In this last case we say that f has a *simple jump discontinuity at* a.)

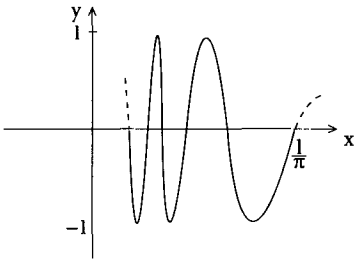

Figure 4.3. Graph of $\sin 1/x$

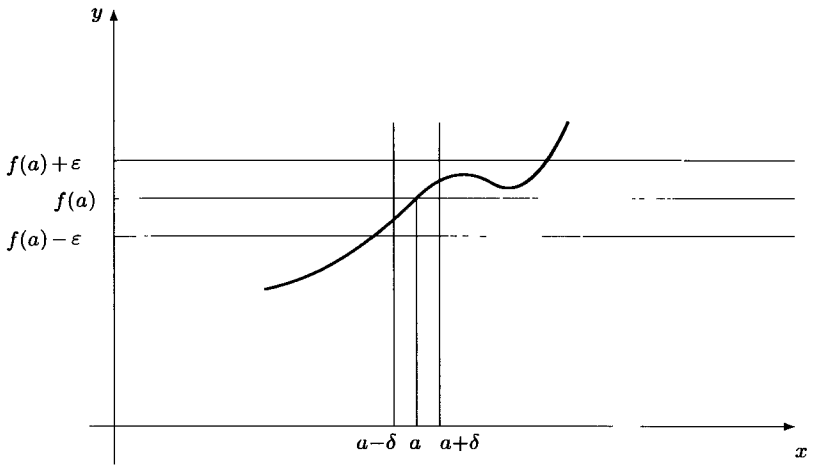

Figure 4.4. Continuity at a

Here are two ways of thinking about continuity of f at a.

(1) In terms of approximations: we can ensure that $f(x)$ approximates $f(a)$ within any prescribed degree of accuracy by choosing x to approximate a sufficiently accurately.

(2) Geometrically: given a horizontal band of any positive width 2ε centred on height $f(a)$, we can choose a vertical band of some suitable width 2δ centred on $x = a$ such that the part of the graph of f in this vertical band is also in the horizontal band (see Figure 4.4): if an aeroplane is flying at 10000 ft at time $t = a$ then it is between 9000 ft and 11000 ft for a non-zero time interval around $t = a$, unless it is capable of discontinuous flight.

The same idea motivates the next result.

Proposition 4.30 *Suppose that* $f : \mathbb{R} \to \mathbb{R}$ *is continous at* $a \in \mathbb{R}$ *and that* $f(a) \neq 0$. *Then there exists* $\delta > 0$ *such that* $f(x) \neq 0$ *whenever* $|x - a| < \delta$.

Proof Take $\varepsilon = |f(a)|/2$ in Definition 4.29. Then there exists $\delta > 0$ such that $|f(x) - f(a)| < |f(a)|/2$ whenever $|x - a| < \delta$. For such x, using the reverse triangle inequality (Corollary 4.10) we get

$$|f(x)| = |f(a)-(f(a)-f(x))| \geqslant |f(a)|-|f(a)-f(x)| > |f(a)|-|f(a)|/2 > 0,$$

so $f(x) \neq 0$. □

In view of such results, continuity is sometimes called 'the principle of inertia'.

As in the definition of $\lim_{x\to a} f(x)$, it is not necessary for f to be defined on all of $\mathbb{R}$ for the definition of continuity of f at a to make sense. It is certainly enough for f to be defined on some open interval I containing a, since then in Definition 4.29 we can take δ small enough so that $x \in I$ whenever $|x-a| < \delta$. Also, we say that f is *continuous at a from the right (or the left)* if $\lim_{x\to a+} f(x)$ (or $\lim_{x\to a-} f(x)$) exists and equals $f(a)$.

Examples of continuous functions

In this section we review how to build up many examples of continuous functions. If $f, g : \mathbb{R} \to \mathbb{R}$ are functions then we can define functions $|f|$, $f+g$, $f.g : \mathbb{R} \to \mathbb{R}$ by the formulae

$|f|(x) = |f(x)|$, $(f+g)(x) = f(x)+g(x)$, $(f.g)(x) = f(x)g(x)$ for all $x \in \mathbb{R}$.

Also, if $\mathcal{Z} = \{x \in \mathbb{R} : g(x) = 0\}$, 'the zero set of g', then we may define $1/g : \mathbb{R} \setminus \mathcal{Z} \to \mathbb{R}$ by $(1/g)(x) = 1/g(x)$ for all $x \in \mathbb{R} \setminus \mathcal{Z}$.

Proposition 4.31 *Suppose that $f, g : \mathbb{R} \to \mathbb{R}$ are continuous at $a \in \mathbb{R}$. Then so are (a) $|f|$, (b) $f+g$ and (c) $f.g$. (d) If $g(a) \neq 0$ then $1/g$ is continuous at a.*

Proof (a) Let $\varepsilon > 0$. We know there exists $\delta > 0$ with $|f(x) - f(a)| < \varepsilon$ whenever $|x-a| < \delta$. Then using the reverse triangle inequality (Corollary 4.10), whenever $|x-a| < \delta$ we have

$$||f|(x) - |f|(a)| = ||f(x)| - |f(a)|| \leqslant |f(x) - f(a)| < \varepsilon$$

so $|f|$ is continuous at a.

(b) For $\varepsilon > 0$ there exists $\delta_1 > 0$ such that $|f(x) - f(a)| < \varepsilon/2$ whenever $|x-a| < \delta_1$, and $\delta_2 > 0$ such that $|g(x) - g(a)| < \varepsilon/2$ whenever $|x-a| < \delta_2$. Let $\delta = \min\{\delta_1, \delta_2\}$. Then whenever $|x-a| < \delta$, we have

$$\begin{aligned}|(f+g)(x) - (f+g)(a)| &= |f(x) - f(a) + g(x) - g(a)| \\ &\leqslant |f(x) - f(a)| + |g(x) - g(a)| < \varepsilon/2 + \varepsilon/2 = \varepsilon.\end{aligned}$$

so $f+g$ is continuous at a.

(c) For the proof that $f.g$ is continuous at $a \in X$ when f and g are, it makes sense to 'begin at the end'. We are going to use a trick way of writing $f(x)g(x) - f(a)g(a)$, as $f(x)(g(x) - g(a)) + (f(x) - f(a))g(a)$ (the roles of f and g could be exchanged). From this we see

$$\begin{aligned}|f(x)g(x) - f(a)g(a)| &= |f(x)(g(x) - g(a)) + (f(x) - f(a))g(a)| \\ &\leqslant |f(x)||g(x) - g(a)| + |f(x) - f(a)||g(a)|.\end{aligned}$$

We know that $|f(x) - f(a)|$ is small when $|x-a|$ is sufficiently small, and $|g(a)|$ is a constant so gives no trouble—given $\varepsilon > 0$ we may choose

$\delta_1 > 0$ such that $|f(x) - f(a)| < \varepsilon/2(|g(a)| + 1)$ whenever $|x - a| < \delta_1$. [The extra 1 is added on the denominator just to avoid making a special case when $g(a) = 0$.] So $|f(x) - f(a)||g(a)| < \varepsilon/2$ whenever $|x - a| < \delta_1$. But $|f(x)||g(x) - g(a)|$ is slightly more awkward to deal with since $|f(x)|$ varies. However, it does not vary too wildly near a since f is continuous at a: there exists $\delta_2 > 0$ such that $|f(x) - f(a)| < 1$ whenever $|x - a| < \delta_2$, so for all such x, we have $|f(x)| = |f(x) - f(a) + f(a)| \leqslant 1 + |f(a)|$ by the triangle inequality. Finally, by continuity of g at a there exists $\delta_3 > 0$ such that $|g(x) - g(a)| < \varepsilon/2(1 + |f(a)|)$ whenever $|x - a| < \delta_3$. Put $\delta = \min\{\delta_1, \delta_2, \delta_3\}$. Then for any x with $|x - a| < \delta$ we have

$$\begin{aligned} |f(x)g(x) - f(a)g(a)| &\leqslant |f(x)||g(x) - g(a)| + |f(x) - f(a)||g(a)| \\ &< \frac{(1 + |f(a)|)\varepsilon}{2(1 + |f(a)|)} + \frac{\varepsilon|g(a)|}{2(|g(a)| + 1)} \\ &< \varepsilon/2 + \varepsilon/2 = \varepsilon. \end{aligned}$$

So $f.g$ is continuous at $a \in X$.

(d) First we note that by continuity of g at a, there is an open interval containing a on which $1/g$ is defined because g is never zero (see Proposition 4.30). Now beginning at the end again, we are going to use

$$\left|\frac{1}{g(x)} - \frac{1}{g(a)}\right| = \frac{|g(a) - g(x)|}{|g(x)||g(a)|}. \qquad (\ddagger)$$

We know that $|g(x) - g(a)|$ is small when $|x - a|$ is small, and $|g(a)|$ is a non-zero constant, so it is easy to handle. But $|g(x)|$ varies, and might 'come dangerously close to 0', so that $\ddagger$ might become large. We get around that as follows. By continuity of g at a, there exists $\delta_1 > 0$ such that $|g(x) - g(a)| < |g(a)|/2$ whenever $|x - a| < \delta_1$. For all such x we have, using the reverse triangle inequality (Corollary 4.10.),

$$|g(x)| = |(g(a) - (g(a) - g(x))| \geqslant |g(a)| - |g(a) - g(x)| \geqslant |g(a)|/2.$$

Continuity of g at a gives $\delta_2 > 0$ such that $|g(x) - g(a)| < \varepsilon|g(a)|^2/2$ whenever $|x - a| < \delta_2$. Put $\delta = \min\{\delta_1, \delta_2\}$. Then using $(\ddagger)$ above, for any x with $|x - a| < \delta$,

$$\left|\frac{1}{g(x)} - \frac{1}{g(a)}\right| = \frac{|g(a) - g(x)|}{|g(x)||g(a)|} \leqslant \frac{|g(x) - g(a)|}{|g(a)|^2/2} < \varepsilon.$$

So $1/g$ is continuous at a. □

The above proofs can be shortened by hiding the secrets of how they are constructed. To illustrate the shorter version, and to see how to reassemble the proof forwards, here is a rabbit-out-of-a-hat proof of (c):

Proof of (c) Let $\varepsilon > 0$. By continuity of f at a, there exists $\delta_1 > 0$ such that $|f(x) - f(a)| < \varepsilon/2(|g(a)| + 1)$ whenever $|x - a| < \delta_1$. Also by continuity of f at a, there exists $\delta_2 > 0$ such that $|f(x) - f(a)| < 1$, and hence $|f(x)| < 1 + |f(a)|$, whenever $|x - a| < \delta_2$. Finally, by continuity of g at a there exists $\delta_3 > 0$ such that $|g(x) - g(a)| < \varepsilon/2(1 + |f(a)|)$ whenever $|x - a| < \delta_3$. Put $\delta = \min\{\delta_1, \delta_2, \delta_3\}$. Then for any x with $|x - a| < \delta$ we have

$$|f(x)g(x) - f(a)g(a)| \leqslant |f(x)||g(x) - g(a)| + |f(x) - f(a)||g(a)| < \varepsilon.$$

□

We can use Proposition 4.31 and induction to see that other real-valued functions are continuous.

Proposition 4.32 *(i) Let $p : \mathbb{R} \to \mathbb{R}$ be the 'polynomial function' defined by $p(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$ where the a_i are constants. Then p is continuous.*

(ii) Let $r : \mathbb{R} \setminus \mathcal{Z} \to \mathbb{R}$ be the rational function $x \mapsto p(x)/q(x)$ where p and q are polynomial functions and $\mathcal{Z}$ is the zero set of q. Then r is continuous on $\mathbb{R} \setminus \mathcal{Z}$.

Proof For (i), it is easy to check that the map $x \mapsto x$ is continuous on $\mathbb{R}$, and so too is any constant function. Next we show inductively that $x \mapsto x^n$ is continuous on $\mathbb{R}$ for any $n \in \mathbb{N}$. The case $n = 1$ is continuity of $x \mapsto x$. Suppose inductively that $x \mapsto x^n$ is continuous on $\mathbb{R}$. Then using (c) of Proposition 4.31, $x \mapsto x.x^n = x^{n+1}$ is continuous on $\mathbb{R}$. Hence by induction, $x \mapsto x^n$ is continuous for any positive integer n. Since the constant map $x \mapsto a_n$ is also continuous on $\mathbb{R}$, another application of (c) shows that $x \mapsto a_n x^n$ is continuous on $\mathbb{R}$. Now an easy induction on (b) shows that any polynomial function is continuous on $\mathbb{R}$. For (ii), an application of (d) shows that $x \mapsto 1/q(x)$ is continuous on $\mathbb{R} \setminus \mathcal{Z}$, and then (c) shows that $x \mapsto p(x)/q(x)$ is continuous on $\mathbb{R} \setminus \mathcal{Z}$. □

Proposition 4.33 *Suppose that $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ are such that f is continuous at $a \in \mathbb{R}$, and g is continuous at $f(a)$. Then $g \circ f$ is continuous at a.*

Proof Let $\varepsilon > 0$. By continuity of g at $f(a)$ there exists $\delta_1 > 0$ such that $|g(y) - g(f(a))| < \varepsilon$ whenever $|y - f(a)| < \delta_1$. By continuity of f at a, there

exists $\delta_2 > 0$ such that $|f(x) - f(a)| < \delta_1$ whenever $|x - a| < \delta_2$. Now for any x with $|x-a| < \delta_2$ we have $|f(x)-f(a)| < \delta_1$ so $|g(f(x))-g(f(a))| < \varepsilon$. This shows that $g \circ f$ is continuous at a. □

Like Proposition 4.31, this result helps build up a store of continuous functions, especially when used in conjunction with continuity of specific functions such as the exponential and log functions, cosine and sine functions, and the like, whose continuity properties we know from analysis (see for example 7.4 of Hart (2001) or Part III of Spivak (2006)). So functions such as $x \mapsto \sin(x^2 + 3x + 1)$, $x \mapsto \mathrm{e}^{-x^2}$, $x \mapsto \mathrm{e}^{x^3+\cos x}$ are continuous on $\mathbb{R}$.

★ Here is a more general approach to continuity for real-valued functions of a real variable.

Definition 4.34 *Let $f : X \to \mathbb{R}$ be a function defined on a subset $X \subseteq \mathbb{R}$ and let $a \in X$. We say f* is continuous at a *if given $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(x) - f(a)| < \varepsilon$ whenever $|x - a| < \delta$ and $x \in X$.*

The more general analogue of Proposition 4.31 can be proved similarly; after each occurrence of the phrase 'whenever $|x - a| < \delta$' we just insert 'and $x \in X$'. This is a special case of the later Proposition 5.17. ★

In connection with the false start we made on defining continuity, the following theorem, usually called the intermediate value theorem, is true.

Theorem 4.35 *Any continuous function $f : \mathbb{R} \to \mathbb{R}$ has the IVP. The same is true for a continuous function $f : I \to \mathbb{R}$ for any interval I in $\mathbb{R}$.*

We could give the proof now, using the completeness property, but before proving this and other basic results about continuity we raise the stakes by generalizing to functions between more general 'spaces' than subsets of $\mathbb{R}$. The motives for this were mentioned in the introduction.

Exercise 4.1 Show that if $\emptyset \neq A \subseteq B \subseteq \mathbb{R}$ and B is bounded above then A is bounded above and $\sup A \leqslant \sup B$

Exercise 4.2 Show that if A and B are non-empty subsets of $\mathbb{R}$ which are bounded above then $A \cup B$ is bounded above and

$$\sup A \cup B = \max\{\sup A, \sup B\}$$

Exercise 4.3 Formulate and prove analogues of Exercises 4.1 and 4 2 for inf.

Exercise 4.4 For each of the following subsets of $\mathbb{R}$ find the sup if it exists, and decide whether it is in the set·

$$\{x : x^2 \leqslant 2x - 1\}. \qquad \{x : x^2 + 2x \leqslant 1\},$$

$$\{x : x^3 < 8\}, \qquad \{x \,.\, x \sin x < 1\}.$$

Exercise 4.5 Show that there is no rational number q such that $q^2 = 2$.
[Hint. express q as a quotient of integers m/n where m, n are mutually prime, and show that $m^2 = 2n^2$ leads to a contradiction.]

Exercise 4.6* Show that if m and n are positive integers with highest common factor 1, then m/n is the square of a rational number if and only if m and n are both squares of integers.

Exercise 4.7 Deduce from the completeness property Proposition 4.4 that a non-empty set of real numbers which is bounded below has a greatest lower bound.

Exercise 4.8 Prove that between any two distinct real numbers there is an irrational number.

Exercise 4.9 Prove that if y, α are real numbers with $y > 1$ then $n^\alpha / y^n \to 0$ as $n \to \infty$.
[Hint: use the binomial expansion of $(1 + x)^n$ where $y = 1 + x$.]

Exercise 4.10 Prove that $\lim_{n\to\infty} n^{1/n} = 1$.
[Hint: Put $n^{1/n} = 1 + a_n$ and note that $a_n > 0$ for $n > 1$ Using $(1 + a_n)^n = n$, deduce that $n - 1 \geqslant n(n-1)a_n^2/2$ for $n > 1$ and hence $0 \leqslant a_n^2 \leqslant 2/n$.]

Exercise 4.11 Given a set of r non-negative real numbers $\{a_1, a_2, \ldots, a_r\}$, let $a = \max\{a_1, a_2, \ldots, a_r\}$. Prove that for any positive integer n,

$$a^n \leqslant a_1^n + a_2^n + \,.\,.\, + a_r^n \leqslant r a^n.$$

By taking nth roots throughout, deduce that

$$a \leqslant (a_1^n + a_2^n + \ldots + a_r^n)^{1/n} \leqslant r^{1/n} a,$$

and hence that $\lim_{n\to\infty} (a_1^n + a_2^n + \,.\,. \, + a_r^n)^{1/n} = a$.

Exercise 4.12 Give an example where $f(x) \to b$ as $x \to a$ and $g(y) \to c$ as $y \to b$ but $g(f(x)) \not\to c$ as $x \to a$

Exercise 4.13 (a) Prove that for any y, z in $\mathbb{R}$,

$$\max\{y, z\} = \frac{1}{2}(y + z + |y - z|), \quad \min\{y, z\} = \frac{1}{2}(y + z - |y - z|).$$

(b) Given that $f, g : \mathbb{R} \to \mathbb{R}$ are continuous at a, prove that h and k are continuous at $a \in \mathbb{R}$, where for any x in $\mathbb{R}$

$$h(x) = \max\{f(x), g(x)\}, \quad k(x) = \min\{f(x), g(x)\}.$$

Exercise 4.14 Let $f, g \cdot \mathbb{R} \setminus \{0\} \to \mathbb{R}$ be given by

$$f(x) = x \sin 1/x, \qquad g(x) = \sin 1/x.$$

Prove that $\lim_{x \to 0} f(x) = 0$, while $\lim_{x \to 0} g(x)$ does not exist.

Exercise 4.15 Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} 0, & x \in \mathbb{Q}, \\ 1, & x \notin \mathbb{Q}. \end{cases}$$

Show that f is not continuous at any point in $\mathbb{R}$.

Exercise 4.16* Let $f : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = \begin{cases} 0 & \text{if } x = 0 \text{ or } x \notin \mathbb{Q}, \\ 1/q & \text{if } x = p/q,\ p, q \text{ integers with highest common factor } 1,\ q > 0. \end{cases}$$

Prove that f is discontinuous at any non-zero a in $\mathbb{Q}$, but continuous at 0 and at any irrational a in $\mathbb{R}$

Exercise 4.17* A function $f : \mathbb{R} \to \mathbb{R}$ is said to be *convex* if

$$f(\lambda x + (1 - \lambda)y) \leqslant \lambda f(x) + (1 - \lambda) f(y)$$

for all x, y in $\mathbb{R}$ and all λ in $[0, 1]$. Prove that any convex function is continuous.

Exercise 4.18** For any function $f : \mathbb{R} \to \mathbb{R}$ show that the set of points $a \in \mathbb{R}$ at which f has a simple jump discontinuity is countable.

5 Metric spaces

In this chapter we begin to study metric spaces. These are a bit more concrete than the topological spaces that we shall study later, but they give valuable pointers for the more abstract material. They are also related to analysis and geometry in intuitively appealing ways.

Motivation and definition

The motivation for metric spaces comes from studying continuity. We begin by rephrasing Definition 4.29 using more English and no Greek: a real-valued function of a real variable is continuous at a if we can make the distance $|f(x) - f(a)|$ between $f(x)$ and $f(a)$ as small as we please by choosing x so that the distance $|x - a|$ between x and a is sufficiently small. (The reader is reminded that the terms function and map are interchangeable. We tend to use the former when dealing with functions of real variables, in which case this terminology is long established, and the latter when dealing with maps between more general sets.) Next let us consider a real-valued function f of two real variables. We again get a definition corresponding to our intuitive idea of continuity by changing the above wording very slightly: f is continuous at a point (a, b) in $\mathbb{R}^2$ if we can make the distance between $f(x, y)$ and $f(a, b)$ as small as we please by choosing (x, y) so that its distance from (a, b) is sufficiently small. We may recover an $\varepsilon - \delta$ form of this definition by using the formulae for the distances involved. Since $f(x, y)$ and $f(a, b)$ are real numbers (f is real-valued) the distance between them is $|f(x, y) - f(a, b)|$. Since (x, y) and (a, b) are points in the plane, the distance between them is $\sqrt{[(x-a)^2 + (y-b)^2]}$, where as always this means the non-negative square root. Thus $f : \mathbb{R}^2 \to \mathbb{R}$ is continuous at (a, b) if given $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(x, y) - f(a, b)| < \varepsilon$ for all (x, y) in $\mathbb{R}^2$ satisfying $\sqrt{[(x-a)^2 + (y-b)^2]} < \delta$.

Now given any positive integer n let us try to define continuity for a real-valued function f of n real variables, $f : \mathbb{R}^n \to \mathbb{R}$. We shall denote a point in $\mathbb{R}^n$ by $x = (x_1, x_2, \ldots, x_n)$. By analogy with our previous definitions we may try writing: f is continuous at $a = (a_1, a_2, \ldots, a_n)$ if the distance between $f(x)$ and $f(a)$ can be made as small as we please by choosing x so that the distance between x and a is sufficiently small. Let

us see if this means anything. Since $f(x)$ and $f(a)$ are real numbers, the distance between them is $|f(x)-f(a)|$. However, continuity for $f : \mathbb{R}^n \to \mathbb{R}$ at a does not make sense (for $n > 3$) until we find a suitable meaning for 'the distance between x and a in $\mathbb{R}^n$' for general n.

Such a meaning is not hard to guess when we look at the particular cases $n = 1$, 2, 3. When $n = 3$ and x and a are points in Euclidean 3-space, the distance between them is $\sqrt{[(x_1 - a_1)^2 + (x_2 - a_2)^2 + (x_3 - a_3)^2]}$. We have already used similar formulae for $n = 1$ and $n = 2$. (For the case $n = 1$ note that $|x_1 - a_1| = \sqrt{(x_1 - a_1)^2}$.) It is therefore plausible to define 'the Euclidean distance between $(x_1, x_2, \ldots, x_n)$ and $(a_1, a_2, \ldots, a_n)$' to be

$$\sqrt{\left[\sum_{i=1}^{n}(x_i - a_i)^2\right]}.$$

Definition 5.1 *A function $f : \mathbb{R}^n \to \mathbb{R}$ is* continuous at a point $a \in \mathbb{R}^n$, say $a = (a_1, a_2, \ldots, a_n)$, *if given $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(x) - f(a)| < \varepsilon$ for every $x = (x_1, x_2, \ldots, x_n)$ satisfying*

$$\sqrt{\left[\sum_{i=1}^{n}(x_i - a_i)^2\right]} < \delta.$$

For general n this does not have any graphical interpretation except by analogy with the cases $n = 1$ and $n = 2$. However, it still has familiar physical interpretations. It often happens that some physical quantity depends on several variables. For example, the energy of a given solid body in a gravitational field depends on its height, its linear velocity, and its angular velocity, and these may be described by seven real variables. Continuity in the sense of Definition 5.1 for the function giving the energy in terms of these seven variables means that if the variables are altered slightly the energy changes only slightly.

Now let us try generalizing one step further. The way in which we arrived at Definition 5.1 suggests a plausible definition of continuity for any map $f : X \to Y$ provided that we can give an adequate meaning to 'the distance between' any two elements or points in X and likewise for any two points in Y. Formally, a function giving the distance between any two points of a set X will be a map $d : X \times X \to \mathbb{R}$, since for any two points x, y of X it should give a real number (the distance between x and y). What properties should this distance function, or *metric* d have? The

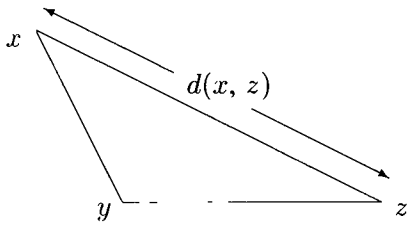

Figure 5.1. Triangle inequality in Euclidean space

choice of these was probably historically a matter of trial and error, but we shall go straight to the historical winners: we pick out some properties of Euclidean distances in the line, plane, and 3-space, and then use them as the *axioms* for a general metric space.

First, the distance between two points is greater than zero except when the points coincide:

(M1) for all $x, y \in X$, $d(x, y) \geqslant 0$; and $d(x, y) = 0$ iff $x = y$.

Secondly, the distance from y to x is the same as the distance from x to y:

(M2) (Symmetry) for all $x, y \in X$, $d(y, x) = d(x, y)$.

Finally we use the *triangle inequality*

(M3) for all $x, y, z \in X$, $d(x, z) \leqslant d(x, y) + d(y, z)$.

Geometrically this says: in any triangle, the length of a side is less than or equal to the sum of the lengths of the other two sides. This is familiar in the plane or in 3-space (see Figure 5.1); in the line, the 'triangle' collapses and we have Proposition 4.9.

It turns out that any function d satisfying just these three properties is similar enough to Euclidean distance for a lot of our geometric intuition about distances to work, so we formalize this into the definition of a metric space.

Definition 5.2 *A* metric space *consists of a non-empty set X together with a function $d : X \times X \to \mathbb{R}$ such that* (M1), (M2), *and* (M3) *above hold.*

We often just talk about 'the metric space X' for short, but there is always a metric d attached to it, which we name only when necessary. The elements of X are called 'points' of the space, and d is called the metric or the distance function. Note that the set X is assumed to be non-empty. The choice between this and allowing the empty set is a matter of swings

and roundabouts- -there are advantages and disadvantages in each. We have chosen the non-empty option because if we allow the empty set, then some later results would need certain spaces to be non-empty in a context where it would be easy to forget to say so.

► At this point readers familiar with vector spaces may refer to the web site for the definition and examples of a concept which lies between Euclidean spaces and metric spaces in degree of generality, that of a *normed vector space*. Any norm on a vector space gives rise to a metric on it. Many of the metric spaces below are actually normed vector spaces. ◄

Before giving examples of metric spaces, we follow the train of thought that led us to them by defining continuity in this context.

Definition 5.3 *Suppose that* (X, d_X) *and* (Y, d_Y) *are metric spaces and let* $f : X \to Y$ *be a map.*

(*a*) *We say* f *is* continuous at $x_0 \in X$ *if given* $\varepsilon > 0$, *there exists* $\delta > 0$ *such that* $d_Y(f(x), f(x_0)) < \varepsilon$ *whenever* $d_X(x, x_0) < \delta$.

(*b*) *We say* f *is* continuous *if* f *is continuous at every* $x_0 \in X$.

When there are other metrics around we say 'f is (d_X, d_Y)-continuous'.

Examples of metric spaces

We shall look at several examples of metric spaces, to get familiar with the definition and to explore its scope. The extent to which metric spaces are a fruitful generalization of Euclidean spaces depends largely on how many interesting examples there are of metric spaces. Some of our examples are designed to illustrate phenomena internal to metric space theory, but others are of interest in analysis. The first example is the one from which we abstracted the definition.

Example 5.4 Euclidean n-space $(\mathbb{R}^n, d_2)$ where for

$$x = (x_1, x_2, \ldots, x_n) \text{ and } y = (y_1, y_2, \ldots, y_n),\ d_2(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

It is easy to see that (M1) and (M2) are satisfied. In order to check (M3), let $z = (z_1, z_2, \ldots, z_n)$, let all summations be over $i = 1, 2, \ldots, n$ and write $x_i - y_i = r_i$, $y_i - z_i = s_i$. Then we have to prove

$$\left(\sum r_i^2\right)^{\frac{1}{2}} + \left(\sum s_i^2\right)^{\frac{1}{2}} \geqslant \left(\sum (r_i + s_i)^2\right)^{\frac{1}{2}}.$$

Since both sides are non-negative, it is equivalent (squaring both sides) to prove

$$\sum r_i^2 + \sum s_i^2 + 2\left(\sum r_i^2\right)^{\frac{1}{2}}\left(\sum s_i^2\right)^{\frac{1}{2}} \geqslant \sum r_i^2 + \sum s_i^2 + 2\sum r_i s_i.$$

This in turn is equivalent to proving Cauchy's inequality,

$$\left(\sum r_i^2\right)\left(\sum s_i^2\right) \geqslant \left(\sum r_i s_i\right)^2.$$

The reader may have seen a proof of this inequality before. One proof is on the web site.

When we consider $\mathbb{R}^n$ as a metric space, this Euclidean metric will always be understood unless some other is specified.

Next we see that from the metric space viewpoint, the complex numbers are very like $(\mathbb{R}^2, d_2)$.

Example 5.5 Let $X = \mathbb{C}$ and for z_1, z_2 in $\mathbb{C}$ let $d(z_1, z_2) = |z_1 - z_2|$. Again (M1) and (M2) clearly hold. When we express each complex number in terms of its real and imaginary parts, the triangle inequality for $\mathbb{C}$, $|z_1 - z_3| \leqslant |z_1 - z_2| + |z_2 - z_3|$ coincides with the triangle inequality for $(\mathbb{R}^2, d_2)$. The close relationship between these metric spaces will be made precise in Example 6.40.

The next example is a much stranger one.

Example 5.6 Let X be any non-empty set and define d by

$$d(x, y) = \begin{cases} 1, & x \neq y, \\ 0, & x = y. \end{cases}$$

It is easy to check that (M1) and (M2) hold. To check (M3), note that if $x = z$ then $d(x, z) = 0$ and certainly $d(x, y) + d(y, z) \geqslant d(x, z)$ since both of $d(x, y)$, $d(y, z)$ are non-negative. If $x \neq z$ then at least one of $x \neq y$, $y \neq z$ must be true, so $d(x, y) + d(y, z)$ is 1 or 2, while $d(x, z) = 1$ so again the triangle inequality holds.

For any set X this metric is called *the discrete metric*. Such 'pathological' examples, as they are nicknamed, are not normally used in applications in analysis. They serve as a warning to check by rigorous proofs that results suggested by intuition really hold in general metric spaces. In

other words, they are potential counterexamples; they explore the boundaries of the concept of a metric space. Nevertheless, metric spaces built up from discrete metric spaces have been used in problems in combinatorics.

The next set of examples show that there are metrics for $\mathbb{R}^n$ other than the Euclidean metric of Example 5.4. We illustrate this in the plane.

Example 5.7 Let $X = \mathbb{R}^2$ and for $x = (x_1, x_2)$, $y = (y_1, y_2)$ let

$$d_1(x, y) = |x_1 - y_1| + |x_2 - y_2|,$$

$$d_2(x, y) = [(x_1 - y_1)^2 + (x_2 - y_2)^2]^{\frac{1}{2}},$$

$$d_\infty(x, y) = \max\{|x_1 - y_1|, |x_2 - y_2|\}.$$

The choice of subscripts is explained in the web site.

We already know that d_2 satisfies the metric space axioms. It is easy to see that d_1 and d_∞ satisfy (M1) and (M2). To check (M3), let $z = (z_1, z_2)$. Then

$$\begin{aligned} d_1(x, y) + d_1(y, z) &= |x_1 - y_1| + |x_2 - y_2| + |y_1 - z_1| + |y_2 - z_2| \\ &= |x_1 - y_1| + |y_1 - z_1| + |x_2 - y_2| + |y_2 - z_2| \\ &\geqslant |x_1 - z_1| + |x_2 - z_2| \\ &= d_1(x, z). \end{aligned}$$

Also, for $i = 1, 2$ we have

$$|x_i - z_i| \leqslant |x_i - y_i| + |y_i - z_i| \leqslant d_\infty(x, y) + d_\infty(y, z),$$

so

$$d_\infty(x, z) \leqslant d_\infty(x, y) + d_\infty(y, z).$$

We could also let d_0 be the discrete metric on $\mathbb{R}^2$. So there may be several distinct metric spaces with the same underlying set. We shall see later that d_1, d_2, d_∞ are all equivalent in a certain sense (and each one deserves to be called a 'product metric'), but that they are not equivalent to d_0.

It is clear how to define analogues of these three metrics on $\mathbb{R}^n$ for any $n \in \mathbb{N}$. The proofs that the axioms hold are similar to the above.

Next we are going to look ways of 'getting new metric spaces from old', which have counterparts for many other mathematical structures. Here they are called (metric) subspaces and products.

Example 5.8 *Metric subspaces.* Suppose that (X, d) is a metric space and that A is a non-empty subset of X. Let $d_A : A \times A \to \mathbb{R}$ be the

restriction of d to $A \times A$ (recall that this means $d_A(x, y) = d(x, y)$ for any x, y in A). The metric space axioms hold for d_A since they hold for d.

The metric space (A, d_A) is called a *(metric) subspace* of (X, d) and d_A is called *the metric on A induced by d*. When it is agreed which metric d is intended we may just say that A is a subspace of X. In this looser terminology we call A either a subset or a subspace of X according to the emphasis desired at the time. If A is a non-empty subset of $\mathbb{R}^n$ then in referring to A as a metric space we assume the metric induced by the Euclidean metric on $\mathbb{R}^n$ unless some other is specified. In particular this applies to subsets of $\mathbb{R}$.

When we have metric spaces (X, d), (X', d'), and a map $f : X \to X'$ then for any subset $A \subseteq X$ and $a \in A$ we can talk about continuity (more precisely, (d_A, d')-continuity) of $f|A$ at a. It is important to distinguish this from continuity (more precisely (d, d')-continuity) of f at a. An extreme example of this is

Example 5.9 Consider $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} 0, & x \in \mathbb{Q}, \\ 1, & x \notin \mathbb{Q}. \end{cases}$$

Then $f|\mathbb{Q} : \mathbb{Q} \to \mathbb{R}$ is the constant function with value 0, and is continuous at every point of $\mathbb{Q}$, whereas f is not continuous at any point. This example might suggest that continuity of $f|A$ is not very useful. However, suppose that a real-valued function f is defined on some subset of $\mathbb{R}$ containing $[a, b]$. Let us see what continuity of $f|[a, b]$ means. At the point a it means: given any $\varepsilon > 0$ there exists $\delta > 0$ such that $|f(x) - f(a)| < \varepsilon$ for all x satisfying $|x - a| < \delta$ *and also* $x \in [a, b]$, or equivalently, for all x satisfying $a \leqslant x < a + \delta$. This means continuity from the right of f at a. Similarly at b it means continuity from the left. Finally, for any $c \in (a, b)$ it means ordinary (two-sided) continuity. Thus continuity of $f|[a, b]$ *is* of interest, as the reader who has studied the mean value theorem in differential calculus knows.

Example 5.10 *Product spaces.* This generalizes Example 5.7. Given two metric spaces (X, d_X) and (Y, d_Y) we can define several metrics on $X \times Y$. For points (x_1, y_1) and $y = (x_2, y_2)$ in $X \times Y$ let

$$\begin{aligned} d_1((x_1, y_1), (x_2, y_2)) &= d_X(x_1, x_2) + d_Y(y_1, y_2), \\ d_2((x_1, y_1), (x_2, y_2)) &= [d_X(x_1, x_2)^2 + d_Y(y_1, y_2)^2]^{\frac{1}{2}}, \\ d_\infty((x_1, y_1), (x_2, y_2)) &= \max\{d_X(x_1, x_2), d_Y(y_1, y_2)\}. \end{aligned}$$

These may be proved to be metrics on $X \times Y$ just as in the case $X = Y = \mathbb{R}$ (see Exercise 5.16). As in Example 5.7 any one of these deserves to be called a product metric. We shall see in the next chapter that they are all equivalent in a certain sense. The definition may be extended to the product of any finite number of metric spaces.

If the only examples of metric spaces were Examples 5.4, 5.6, 5.7, together with metric subspaces and products formed from them, it is doubtful whether general metric space theory would be worthwhile. The examples below indicate the wide range of metric space theory (but do not exhaust it). First we sketch examples arising in number theory and group theory, respectively.

Example 5.11 Let p be a fixed prime number, and define $d : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$ by $d(m, m) = 0$ and for $m \neq n$, $d(m, n) = 1/r$ where p^{r-1} is the highest power of p which divides $m - n$.

(M1) and (M2) are easy to check. For (M3), suppose that $m - n = p^{r-1}k$ and $n - q = p^{s-1}k'$, where k and k' are not divisible by p. We can check that $m - q = p^{t-1}k''$ where $t \geqslant \min\{r, s\}$ (equality holds when $r \neq s$) and k'' is not divisible by p. So

$$\begin{aligned} d(m, q) = 1/t &\leqslant 1/(\min\{r, s\}) = \max\{1/r, 1/s\} \\ &= \max\{d(m, n), d(n, q)\} \leqslant d(m, n) + d(n, q). \end{aligned}$$

Example 5.12 ▶ This example will make sense only if you know about groups and generating sets. Suppose G is a finitely generated group and $\mathcal{A}$ is a generating set for G. Let $F(\mathcal{A})$ be the free group on $\mathcal{A}$, and let $p : F(\mathcal{A}) \to G$ be the natural (onto) map. *The word metric $d_{\mathcal{A}}$ on G associated to $\mathcal{A}$* is defined as follows: for $g_1, g_2 \in G$ let $d_{\mathcal{A}}(g_1, g_2)$ be the length of the shortest word in $p^{-1}(g_1^{-1}g_2)$.

Again (M1) and (M2) are easy to check. For (M3), if w is a word of shortest length in $p^{-1}(g_1^{-1}g_2)$ and w' is a word of shortest length in $p^{-1}(g_2^{-1}g_3)$ then ww' is a word in $p^{-1}(g_1^{-1}g_3)$, and

$$\begin{aligned} d_{\mathcal{A}}(g_1, g_3) \leqslant \text{ length } (ww') &= \text{ length}(w) + \text{ length}(w') \\ &= d_{\mathcal{A}}(g_1, g_2) + d_{\mathcal{A}}(g_2, g_3). \end{aligned}$$ ◀

There are two further kinds of metric spaces used in analysis: sequence spaces and function spaces. We discuss sequence spaces on the web site, and introduce function spaces here. We take some collection of functions and decide to treat it as a 'space', calling the individual functions 'points' and putting a metric on the collection.

To indicate how this might be useful, let us consider a classical problem in one of the first areas where function space language was seen to be appropriate, calculus of variations. The brachistocrone problem is roughly as follows. Suppose we have any two points x, y in a vertical plane, with x higher than, but not vertically above, y. What is the shape of the curve in this plane along which a heavy particle will take the least time to slide from x to y under the action of gravity (with no friction)? It is not important for this illustration to be precise about what kind of curves are intended, but we could take 'curve' to mean one defined by a function $h : [0, 1] \to \mathbb{R}^2$ given by $h(t) = (f(t), g(t))$ where $f, g : [0, 1] \to \mathbb{R}$ are continuously differentiable functions. For any given curve λ we can use integration to calculate the time $T(\lambda)$ the particle takes to make the journey from x to y along λ. Thus we get a real-valued function T defined on the collection of all curves from x to y, and the brachistochrone problem is to find the 'point' λ_0 (if there is one) at which T takes a minimum value. (The answer turns out to be part of a cycloid.) This is like looking for the minimum of a function $T : \mathbb{R} \to \mathbb{R}$, except that the domain $\mathbb{R}$ is replaced by the 'space' of curves. The calculus of variations develops analogues of ordinary calculus for solving such problems. Even to begin calculus, we need continuity of T, and this motivates putting a metric on the collection of curves. (The collection of curves is more or less the set of above functions such as h, satisfying $h(0) = x$, $h(1) = y$, except that distinct functions may define the same geometric curve for example, h_1, $h_2 : [0, 1] \to \mathbb{R}$ given by $h_1(t) = (t, t)$ and $h_2(t) = (t^2, t^2)$ both describe a straight line segment joining the points $(0, 0)$ and $(1, 1)$.)

In this and other contexts where maps are defined on collections of functions, the language of function spaces is useful. The possibilities it allows for the use of geometric intuition have proved to be fruitful. We now give a few examples of function spaces with metrics on them.

Example 5.13 Let X be the set of all bounded functions $f : [a, b] \to \mathbb{R}$. Given two points f and g in X, let

$$d(f, g) = \sup_{x \in [a, b]} |f(x) - g(x)|.$$

The right-hand side exists, since f and g are bounded, so there are constants K, L such that $|f(x)| \leqslant K$, $|g(x)| \leqslant L$ for all $x \in [a, b]$ and we have

$$|f(x) - g(x)| \leqslant |f(x)| + |g(x)| \leqslant K + L \quad \text{for all } x \in [a, b].$$

We shall now check in detail that the metric space axioms hold.

(M1) It is clear that $d(f, g) \geqslant 0$ since it is the sup of a (bounded, non-empty) set of non-negative real numbers. Also, if f and g are the same point in X, this means they are identical as functions from $[a, b]$ to $\mathbb{R}$, so $f(x) = g(x)$ for all $x \in [a, b]$, and from its definition $d(f, g) = 0$. Finally, $d(f. g) = 0$ says that $\sup_{x \in [a, b]} |f(x) - g(x)| = 0$, so $f(x) = g(x)$ for all $x \in [a, b]$ which says that $f = g$.

(M2) For any $f, g \in X$ we have $|f(x) - g(x)| = |g(x) - f(x)|$ for all $x \in [a. b]$ so
$\sup_{x \in [a, b]} |f(x) - g(x)| = \sup_{x \in [a, b]} |g(x) - f(x)|$, which says $d(f, g) = d(g, f)$.

(M3) Let $f, g, h \in X$. For any $c \in [a. b]$,

$$\begin{aligned} |f(c) - h(c)| &\leqslant |f(c) - g(c)| + |g(c) - h(c)| \\ &\leqslant \sup_{x \in [a, b]} |f(x) - g(x)| + \sup_{x \in [a, b]} |g(x) - h(x)| \\ &= d(f, g) + d(g, h), \end{aligned}$$

where the first inequality is just the triangle inequality in $\mathbb{R}$. The above holds for any $c \in [a, b]$, so $d(f, g) + d(g, h)$ is an upper bound for the set

$$S = \{|f(c) - h(c)| : c \in [a, b]\}.$$

Hence $d(f, g) + d(g, h) \geqslant \sup S = d(f, h)$ as required.

This metric is called the *sup metric* or the *uniform metric*. We denote the resulting metric space by $(\mathcal{B}([a, b], \mathbb{R}), d_\infty)$, but note that this notation is not universally agreed.

Any continuous function $f : [a, b] \to \mathbb{R}$ is bounded, by a theorem quoted in the introduction, so the set of all such continuous functions forms a subspace of $\mathcal{B}([a, b], \mathbb{R})$, sometimes written $C[a, b]$.

Example 5.14 Let X be the set of all continuous functions $f : [a, b] \to \mathbb{R}$ but this time let

$$d(f, g) = \int_a^b |f(t) - g(t)| \, \mathrm{d}t.$$

To check the metric space axioms we need the following lemma from integration theory.

Lemma 5.15 *Suppose that $h : [a, b] \to \mathbb{R}$ is continuous, that $h(t) \geqslant 0$ for all $t \in [a, b]$ and that $\int_a^b h(t) \, \mathrm{d}t = 0$. Then $h(t) = 0$ for all $t \in [a, b]$.*

The idea of the proof is that if $h(c) > 0$ for some $c \in [a, b]$ then by continuity $h(t)$ exceeds some fixed positive number, for example $\frac{1}{2}h(c)$, throughout an interval of non-zero length around c. This makes a strictly positive contribution to the integral which cannot be cancelled out elsewhere since $h(t)$ is never negative.

(M1) It is clear that $d(f, g) \geqslant 0$ for all $f, g \in X$, and that if $f = g$ then $d(f, g) = 0$. If $d(f, g) = 0$ then by Lemma 5.15 applied with $h = |f - g|$ we get $f = g$.

(M2) Symmetry of $d(f, g)$ is clear.

(M3) For any continuous $f, g, h : [a, b] \to \mathbb{R}$ and any $t \in [a, b]$,

$$|f(t) - h(t)| \leqslant |f(t) - g(t)| + |g(t) - h(t)|.$$

Hence by integration theory,

$$\int_a^b |f(t) - h(t)|\,\mathrm{d}t \leqslant \int_a^b |f(t) - g(t)|\,\mathrm{d}t + \int_a^b |g(t) - h(t)|\,\mathrm{d}t,$$

as required. This metric is called the L^1 metric and sometimes written d_1.

Examples 5.13 and 5.14 give us a choice of two metrics, d_∞ and d_1, on the set of continuous functions $f : [a, b] \to \mathbb{R}$. The metric used in any particular situation depends on which properties of the functions are of interest at the time. When we regard g as a good approximation to f iff $g(t)$ is uniformly close to $f(t)$ for all $t \in [a, b]$, we use d_∞ (this will be studied in Chapter 16). On the other hand, we might not be as much interested in the difference in values of the functions at each point as in their average deviation from one another over the range $[a, b]$. We might then use d_1 or some other metric involving integration, such as in the next example.

Example 5.16 Let X be as in Example 5.14 and let

$$d_2(f, g) = \left\{ \int_a^b (f(t) - g(t))^2\,\mathrm{d}t \right\}^{\frac{1}{2}}.$$

Again (M1) follows from Lemma 5.15, and (M2) clearly holds. The proof that (M3) holds is similar to the proof in Example 5.4 with Cauchy's inequality replaced by its analogue for integrals, the Cauchy–Schwarz inequality:

$$\int_a^b (f(t))^2\,\mathrm{d}t \int_a^b (g(t))^2\,\mathrm{d}t \geqslant \left\{ \int_a^b f(t)g(t)\,\mathrm{d}t \right\}^2,$$

which is proved on the companion web site. The metric d_2 is called the L^2 metric.

Results about continuous functions on metric spaces

Here is a generalization of Proposition 4.31. If $f, g : X \to \mathbb{R}$ are real-valued functions on a metric space X then we can define associated functions $|f|, f+g, f.g : X \to \mathbb{R}$ where, for all $x \in X$,

$$|f|(x) = |f(x)|, \quad (f+g)(x) = f(x) + g(x), \quad (f.g)(x) = f(x)g(x).$$

Also, if g never takes the value 0 on X then we may define $1/g : X \to \mathbb{R}$ by $(1/g)(x) = 1/g(x)$ for all $x \in X$.

Proposition 5.17 *Suppose that $f, g : X \to \mathbb{R}$ are continuous real-valued functions on a metric space (X, d). Then so are (a) $|f|$, (b) $f+g$, and (c) $f.g$. (d) Also, if g is never zero on X, then $1/g$ is continuous on X.*

Proof Let d be the metric on X. Then in (a) (d) continuity at any point $a \in X$ can be proved by an exact replica of the proof of Proposition 4.31: we simply replace the domain $\mathbb{R}$ of the functions by X and every occurrence of '$|x-a| < \delta$' by '$d(x, a) < \delta$'. □

An alternative proof will be given shortly.

The next four results will be generalized in Chapter 8.

Proposition 5.18 *Suppose that $f : X \to Y$ and $g : Y \to Z$ are maps of metric spaces with metrics d_X, d_Y, d_Z, that f is continuous at $a \in X$ and that g is continuous at $f(a)$. Then $g \circ f$ is continuous at a.*

Proof Let $\varepsilon > 0$. Since g is continuous at $f(a)$ there exists $\delta_1 > 0$ such that $d_Z(g(y), g(f(a))) < \varepsilon$ whenever $d_Y(y, f(a)) < \delta_1$, and then by continuity of f at a, there exists $\delta_2 > 0$ such that $d_Y(f(x), f(a)) < \delta_1$ whenever $d_X(x, a) < \delta_2$. Then whenever $d_X(x, a) < \delta_2$ we have $d_Y(f(x), f(a)) < \delta_1$ so $d_Z(g(f(x)), g(f(a))) < \varepsilon$. This gives continuity of $g \circ f$ at a. □

The next three results involve product metric spaces. As we have already mentioned, we shall see in the next chapter that the metrics in Example 5.10 are all equivalent in a sense which means that using any one of them would make the next three propositions true. But in the meantime we shall use the metric called d_1 in Example 5.10 whenever we consider a product of metric spaces.

Proposition 5.19 *Suppose that $f : X \to X'$, $g : Y \to Y'$ are maps of metric spaces which are continuous at $a \in X$, $b \in Y$ respectively. Then the map $f \times g : X \times Y \to X' \times Y'$ given by $(f \times g)(x, y) = (f(x), g(y))$, for all $(x, y) \in X \times Y$, is continuous at (a, b).*

Proof Let d_X, d_Y, $d_{X'}$, $d_{Y'}$ be the metrics on X, Y, X', Y'. Recall we are using the metrics d_1, d_1' on $X \times Y$, $X' \times Y'$, where $d_1((x_1, y_1), (x_2, y_2))$ is defined to be $d_X(x_1, x_2) + d_Y(y_1, y_2)$ and similarly for d_1' (see Example 5.10).

Let $\varepsilon > 0$. It follows from continuity of f at a and g at b that there exist $\delta_1 > 0$, $\delta_2 > 0$ such that $d_{X'}(f(x), f(a)) < \varepsilon/2$ whenever $d_X(x, a) < \delta_1$, and $d_{Y'}(g(y), g(b)) < \varepsilon/2$ whenever $d_Y(y, b) < \delta_2$. Put $\delta = \min\{\delta_1, \delta_2\}$. If $d_1((x, y), (a, b)) < \delta$ then $d_X(x, a) \leqslant d_1((x, y), (a, b)) < \delta \leqslant \delta_1$ and similarly $d_Y(y, b) < \delta_2$ so

$$d_1'((f(x), g(y)), (f(a), g(b))) = d_{X'}(f(x), f(a)) + d_{Y'}(g(y), g(b)) < \varepsilon.$$

This proves that $f \times g$ is continuous at (a, b). □

Proposition 5.20 *The projections $p_X : X \times Y \to X$ $p_Y : X \times Y \to Y$ of a metric product onto its factors, defined by $p_X(x, y) = x$, $p_Y(x, y) = y$, are continuous.*

Proof We again use the metric d_1 on $X \times Y$ as in the proof of Proposition 5.19. We check continuity of p_X at $(a, b) \in X \times Y$. Let $\varepsilon > 0$ and choose $\delta = \varepsilon$. Then whenever $d_1((x, y), (a, b)) < \delta$ we have

$$d_X(p_X(x, y), p_X(a, b)) = d_X(x, a) \leqslant d_1((x, y), (a, b)) < \delta = \varepsilon,$$

so p_X is continuous at (a, b), and similarly for p_Y. □

Definition 5.21 *The* diagonal map $\Delta : X \to X \times X$ *of any set X is the map defined by $\Delta(x) = (x, x)$.*

Proposition 5.22 *The diagonal map $\Delta : X \to X \times X$ of any metric space X is continuous.*

Proof As before we use the metric d_1 on $X \times X$ defined by

$$d_1((x_1, x_2), (x_1', x_2')) = d_X(x_1, x_1') + d_X(x_2, x_2').$$

Let $\varepsilon > 0$. Put $\delta = \varepsilon/2$. Then whenever $d_X(x, x') < \delta$ we have

$$d_1(\Delta(x), \Delta(x')) = d_1((x, x), (x', x')) = d_X(x, x') + d_X(x, x') < \varepsilon.$$

This establishes continuity of Δ. □

We can use these results to give a slightly different proof of Proposition 5.17. Note that as special cases of Proposition 4.31 the functions $\mathbb{R} \to \mathbb{R}$ given by $x \mapsto |x|$ and $\mathbb{R} \setminus \{0\} \to \mathbb{R}$ given by $x \mapsto 1/x$ are

both continuous. Hence if $f, g : X \to \mathbb{R}$ are continuous real-valued functions on a metric space X with $g(x)$ never 0, then by Proposition 5.18 the compositions $x \mapsto f(x) \mapsto |f(x)|$ and $x \mapsto g(x) \mapsto 1/g(x)$ are continuous.

Next, the projections $p_1, p_2 : \mathbb{R}^2 = \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ of $\mathbb{R}^2$ onto the coordinate axes are continuous as a special case of Proposition 5.20, hence by Proposition 4.31 so is their sum $(x, y) \mapsto x + y$ and their product $(x, y) \mapsto xy$. Now suppose $f, g : X \to \mathbb{R}$ are real-valued maps on a metric space X which are continuous at $a \in X$. The sum/product of f and g is the composition

$$X \xrightarrow{\Delta} X \times X \xrightarrow{f \times g} \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R},$$

where the third map is either $(x, y) \mapsto x + y$ or $(x, y) \mapsto xy$. For the composition is $x \mapsto (x, x) \mapsto (f(x), g(x)) \mapsto f(x) + g(x)$ or $f(x)g(x)$. By the above results this composition is continuous at a.

Bounded sets in metric spaces

One topic in metric spaces which intuition guides us to generalize easily from Euclidean spaces is that of bounded sets.

Definition 5.23 *A subset S of a metric space (X, d) is* bounded *if there exist $x_0 \in X$ and $K \in \mathbb{R}$ such that $d(x, x_0) \leqslant K$ for all $x \in S$.*

If S satisfies the definition for some $x_0 \in X$ and $K \in \mathbb{R}$, then it also satisfies the definition with x_0 replaced by any other point $x_1 \in X$ and K replaced by $K + d(x_0, x_1)$. For if $d(x, x_0) \leqslant K$ then

$$d(x, x_1) \leqslant d(x, x_0) + d(x_0, x_1) \leqslant K + d(x_0, x_1).$$

If S satisfies 5.23 then $d(x, y) \leqslant d(x, x_0) + d(x_0, y) \leqslant 2K$ for all $x, y \in S$. The following definition therefore makes sense.

Definition 5.24 *If S is a non-empty bounded subset of a metric space with metric d, then the* diameter of S *is* $\sup\{d(x, y) : x, y \in S\}$. *The diameter of the empty set is* 0.

Definition 5.25 *If $f : S \to X$ is a map from a set S to a metric space X, then we say f* is bounded *if the subset $f(S)$ of X is bounded.*

Here is a sample of the kind of result that our intuition about bounded sets in Euclidean spaces suggests.

Proposition 5.26 *The union of any finite number of bounded subsets of a metric space is bounded.*

Proof It is enough to prove this for two bounded sets, since the result then follows by induction. Before reading on, try to think how the proof should work by contemplating two bounded sets in the plane say. Suppose that S_1, S_2 are bounded subsets of a metric space X with metric d. Then there exist points $x_1, x_2 \in X$ and real numbers K_1, K_2 such that $d(x, x_1) \leqslant K_1$ for all $x \in S_1$ and $d(x, x_2) \leqslant K_2$ for all $x \in S_2$. Put $K = \max\{K_1,\ K_2 + d(x_2, x_1)\}$. Then for any $x \in S_1 \cup S_2$ we have either $x \in S_1$, so $d(x, x_1) \leqslant K_1 \leqslant K$, or else $x \in S_2$, in which case $d(x, x_1) \leqslant d(x, x_2) + d(x_2, x_1) \leqslant K_2 + d(x_2, x_1) \leqslant K$. This shows that $S_1 \cup S_2$ is bounded. □

Open balls in metric spaces

In this section we develop some terminology which is useful for discussing continuity in metric spaces, and will lead us towards a more general framework in which to discuss continuity.

Definition 5.27 *Let (X, d) be a metric space, $x_0 \in X$, and $r > 0$ a real number. The* open ball in X of radius r centred on x_0 *is the set*

$$B_r(x_0) = \{x \in X : d(x, x_0) < r\}.$$

If we are considering more than one metric on X then we write $B_r^d(x_0)$.

Both name and notation vary. Sometimes it is called an 'open spherical neighbourhood'. Notation: we are using B for 'ball'; some others use D for 'disc'.

Example 5.28 (a) In $\mathbb{R}$ (with its usual metric) $B_r(x_0)$ is the open interval $(x_0 - r,\ x_0 + r)$.
(b) Let $X = \mathbb{R}^2$ and $d = d_2$, the Euclidean metric. Then $B_r(x_0)$ is the open disc of radius r centred on x_0 (the set of all points strictly inside the circle of radius r centred on x_0).
(c) Let $X = \mathbb{R}^3$, $d = d_2$. Then $B_r(x_0)$ is the open ball of radius r centred on x_0 (the set of all points strictly inside the sphere of radius r centred on x_0).
(d) Let $X = \mathbb{R}^2$, $d = d_1$ (see Example 5.7). Then $B_r(x_0)$ is the inside of the square centred on x_0 with diagonals of length $2r$ parallel to the axes, as in Figure 5.2.

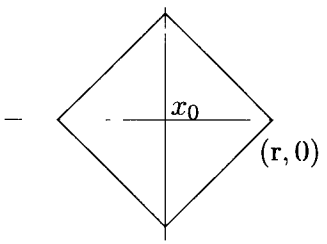

Figure 5.2. $B_r^{d_1}(x_0)$

(e) Let $X = \mathbb{R}^2$ and let d be the discrete metric. Then

$$B_r^d(x_0) = \begin{cases} \{x_0\} & \text{if } r \leqslant 1, \\ \mathbb{R}^2 & \text{if } r > 1. \end{cases}$$

(f) Let X be the set $\mathcal{B}([0, 1], \mathbb{R})$ of all bounded real-valued functions on $[0, 1]$, and let d be the sup metric d_∞. Then for $f_0 \in X$ and $r > 0$ a real number, $B_r^d(f_0)$ is the set of all functions $f \in X$ whose graphs lie inside a ribbon of vertical width $2r$ centred on the graph of f_0 (see Figure 5.3).

Examples 5.28(d) (e) warn us not to take the name 'ball' too seriously balls are not always round. Examples 5.28(b), (d), (e) show that $B_r^d(x_0)$ depends in general on d. It also depends on the underlying set in the way shown by the next example.

Example 5.29 Let $A = [0, 1] \subseteq \mathbb{R}$ with the Euclidean metric d on $\mathbb{R}$ and the induced metric d_A on A. Then we have $B_1^d(1) = (0, 2)$ while on the other hand $B_1^{d_A}(1) = (0, 1]$.

We mention two things that can be done with open balls before going on. First, we may rephrase the definition of a bounded set: a subset S of

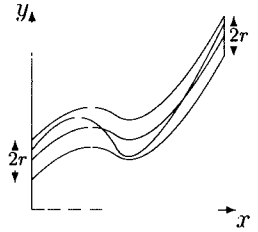

Figure 5.3. Open ball in $(\mathcal{B}([0, 1], \mathbb{R}), d_\infty)$

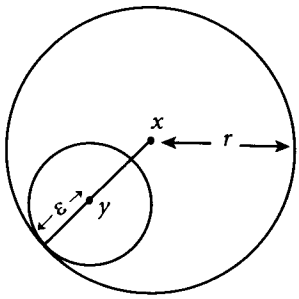

Figure 5.4. Proof of Proposition 5.3

a metric space X is bounded iff $S \subseteq B_r(x_0)$ for some $x_0 \in X$ and $r > 0$. Secondly, we can re-express the definition of continuity in terms of open balls.

Proposition 5.30 *With notation as in Definition 5.3, f is continuous at x_0 iff given $\varepsilon > 0$ there exists $\delta > 0$ such that $f(B_\delta^{d_X}(x_0)) \subseteq B_\varepsilon^{d_Y}(f(x_0))$.*

Proof This is an immediate translation of Definition 5.3. □

We end this section with an important property of open balls whose proof illustrates how geometric intuition and analytic rigour both play a role in metric space theory.

Proposition 5.31 *Given an open ball $B_r(x)$ in a metric space (X, d) and a point $y \in B_r(x)$, there exists $\varepsilon > 0$ such that $B_\varepsilon(y) \subseteq B_r(x)$.*

Proof In the plane, this asserts that we can draw a disc around y lying entirely within the larger disc in Figure 5.4. This is obvious in the picture, but the proof for general metric spaces will have to use the axioms only. However, the picture helps by suggesting what size to try taking ε, namely such that $\varepsilon + d(y, x) \leqslant r$.

Here is the formal proof. Take $\varepsilon = r - d(y, x)$ We note that then $\varepsilon > 0$ since $y \in B_r(x)$ so $d(y, x) < r$. We shall prove that $B_\varepsilon(y) \subseteq B_r(x)$. For if $z \in B_\varepsilon(y)$ then $d(z, y) < \varepsilon$, so $d(z, x) \leqslant d(z, y) + d(y, x) < \varepsilon + d(y, x) = r$, and $z \in B_r(x)$ as required. □

Open sets in metric spaces

Despite the usefulness of open balls, we want a similar but more widely applicable concept generalizing them. Specifically, we generalize the property of open balls expressed in Proposition 5.31, which has been described as 'having some elbow-room around each point'.

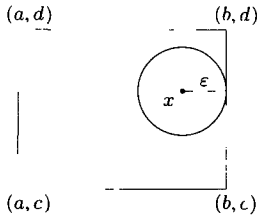

Figure 5.5. Open rectangle

Definition 5.32 *Let (X, d) be a metric space and $U \subseteq X$. We say that U is* open in X *if for every $x \in U$ there exists $\varepsilon_x > 0$ such that $B_{\varepsilon_x}(x) \subseteq U$.*

We have put a suffix x on the ε here to emphasize that in general the size of ε that does the trick will depend on the position of x in the set U. But hereafter we nearly always leave off this x—it is understood that ε depends on x, but the notation becomes clumsy if we insist on emphasising it.

Example 5.33 By Proposition 5.31, any open ball in a metric space X is open in X. In particular any 'open interval' in $\mathbb{R}$ is open in $\mathbb{R}$. On the other hand, intervals in $\mathbb{R}$ such as $[a, b]$, $[a, b)$, $(a, b]$ are not open in $\mathbb{R}$: for $a \in [a, b)$, but no matter how small a positive ε we choose, $B_\varepsilon(a)$ contains points, such as $a - \varepsilon/2$, to the left of a, which are not in $[a, b)$. Note that not every open set is an open ball: for example, in $\mathbb{R}^2$ let U be the interior of a rectangle, say

$$U = \{(x_1, x_2) \in \mathbb{R}^2 : a < x_1 < b,\ c < x_2 < d\}.$$

If $x = (x_1, x_2) \in U$ and we set $\varepsilon = \min\{x_1 - a,\ b - x_1,\ x_2 - c,\ d - x_2\}$, it is easily seen (compare Figure 5.5) that $B_\varepsilon(x) \subseteq U$.

As these Euclidean examples suggest, there are no 'boundary points' in a set U which is open in a metric space—from any point in U one can 'go' some positive distance without going outside U—each point in U has some elbow-room around it, within U.

Example 5.34 For any metric space X, the whole set X and the empty set $\emptyset$ are both open in X. This follows trivially from the definition of 'open'. For example $[a, b]$ is open in $[a, b]$.

Example 5.35 In a discrete metric space X, *any* subset $A \subseteq X$ is open in X. For if $x \in A$ we can choose ε_x to be 1 say, and then $B_{\varepsilon_x}(x) = \{x\} \subseteq A$.

The next examples show that when we say a set is open we have to be careful about which metric space we mean, both about the underlying set and also about the metric.

Example 5.36 A singleton set such as $\{0\}$ is open in $\mathbb{R}$ with the discrete metric, but not in $\mathbb{R}$ with the usual (Euclidean) metric. When necessary we say that a set is 'd-open'. The interval $[a, b]$ is open in $[a, b]$ with its usual metric, as in Example 5.34 above, but not in the larger space $\mathbb{R}$. The interval (a, b) is open in $\mathbb{R}$, but not in $\mathbb{R}^2$ when we identify (a, b) with $(a, b) \times \{0\}$): for $x \in (a, b) \times \{0\}$ there is no $\varepsilon > 0$ such that the disc $B_\varepsilon(x)$ in $\mathbb{R}^2$ is contained in $(a, b) \times \{0\}$—any such disc contains points which are off the x_1-axis see Figure 5.6.

Next we derive yet another criterion for a map between metric spaces to be continuous, this time in terms of open sets. The reader may think we are not making much progress, but merely juggling with definitions. This is true, but eventually this criterion will lead us to generalizing our whole framework to topological spaces. The criterion says that everything about continuity in metric spaces is entirely encoded in the open sets of the spaces: if we know what the open sets in the spaces are, then a function from one metric space to another is continuous iff the inverse image of any open set is open.

If you feel at all shaky about inverse images of sets, before reading the next definition would be a good time to study Chapter 3.

Proposition 5.37 *Suppose that $f : X \to Y$ is a map of metric spaces. Then f is continuous iff $f^{-1}(U)$ is open in X whenever U is open in Y.*

Proof First suppose that f is continuous and that $U \subseteq Y$ is open in Y. We want to show that $f^{-1}(U)$ is open in X. So let $x_0 \in f^{-1}(U)$. Then $f(x_0) \in U$, and since U is open in Y there exists $\varepsilon > 0$ such that $B_\varepsilon(f(x_0)) \subseteq U$. Since f is continuous at x_0, there exists $\delta > 0$ such that $f(B_\delta(x_0)) \subseteq B_\varepsilon(f(x_0))$. From this we get $f(B_\delta(x_0)) \subseteq U$, so $B_\delta(x_0) \subseteq f^{-1}(U)$ and $f^{-1}(U)$ is open in X as required.

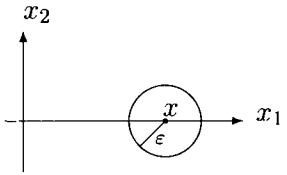

Figure 5.6. $(a, b) \times \{0\}$ not open in $\mathbb{R}^2$

Conversely suppose that $f^{-1}(U)$ is open in X whenever U is open in Y. We shall prove that f is continuous at any $x_0 \in X$. For let $\varepsilon > 0$. Then $B_\varepsilon(f(x_0))$ is open in Y by Proposition 5.31 so $f^{-1}(B_\varepsilon(f(x_0)))$ is open in X. Also, $x_0 \in f^{-1}(B_\varepsilon(f(x_0)))$ since $f(x_0) \in B_\varepsilon(f(x_0))$. So there exists $\delta > 0$ such that $B_\delta(x_0) \subseteq f^{-1}(B_\varepsilon(f(x_0)))$. Then $f(B_\delta(x_0)) \subseteq B_\varepsilon(f(x_0))$, and f is continuous at x_0 by Proposition 5.30. □

Example 5.38 The reader should be warned that when $f : X \to Y$ is a continuous map of metric spaces, it is not necessarily true that the *forwards* image of an open set is open, that is to say, U may be open in X without $f(U)$ being open in Y. For example if we let $f : \mathbb{R} \to \mathbb{R}$ be a constant map, say $f(x) = 0$ for all $x \in \mathbb{R}$, then certainly f is continuous, but for example $(0, 1)$ is open in $\mathbb{R}$ while $f((0, 1)) = \{0\}$ is not open in $\mathbb{R}$.

We end this chapter with two results which show that 'open set' is a more flexible concept than 'open ball'. They feature again later.

Proposition 5.39 *If $U_1, U_2, \ldots, U_m$ are open in a metric space X then so is $\bigcap_{i=1}^{m} U_i$.*

Proof Let $x \in \bigcap_{i=1}^{m} U_i$. Then $x \in U_i$ for each $i = 1, 2, \ldots, m$, so there exists $\varepsilon_i > 0$ such that $B_{\varepsilon_i}(x) \subseteq U_i$. Put $\varepsilon = \min\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m\}$. Then

$$B_\varepsilon(x) \subseteq B_{\varepsilon_i}(x) \subseteq U_i \text{ for each } i = 1, 2, \ldots, m, \text{ so } B_\varepsilon(x) \subseteq \bigcap_{i=1}^{m} U_i$$

and $\bigcap_{i=1}^{m} U_i$ is open as required. □

Thus the intersection of a finite number of open sets is open. Without finiteness, the result is false in general.

Example 5.40 In $\mathbb{R}$, the interval $(-1/n, 1/n)$ is open for each $n \in \mathbb{N}$. But $\bigcap_{n=1}^{\infty} (-1/n, 1/n) = \{0\}$. To see this, note that $0 \in (-1/n, 1/n)$ for every $n \in \mathbb{N}$, so $0 \in \bigcap_{n=1}^{\infty} (-1/n, 1/n)$. On the other hand if $x \neq 0$ then x is not in this intersection, since for sufficiently large n, $x \notin (-1/n, 1/n)$. Since $\{0\}$ is not open in $\mathbb{R}$, we have the required example.

Proposition 5.41 *The union of any collection of sets open in a metric space X is open in X.*

Proof Let I be an indexing set, and for each $i \in I$ let U_i be an open subset of the metric space X. We shall show that $\bigcup_{i\in I} U_i$ is open in X. Let $x \in \bigcup_{i\in I} U_i$. We have $x \in U_{i_0}$ for some $i_0 \in I$, so there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq U_{i_0}$. Then $B_\varepsilon(x) \subseteq \bigcup_{i\in I} U_i$, and the latter is open in X. □

We note that in general neither an intersection nor a union of open balls is again an open ball: this illustrates the greater flexibility of open sets.

Exercise 5.1 Given points x, y, z in a metric space (X, d) prove that

$$|d(x, z) - d(y, z)| \leqslant d(x, y)$$

Exercise 5.2 Given points x, y, z, t in a metric space (X, d) prove that

$$|d(x, y) - d(z, t)| \leqslant d(x, z) + d(y, t)$$

Exercise 5.3 Given points $x_1, x_2, \ldots, x_n$ in a metric space (X, d) prove that

$$d(x_1, x_n) \leqslant d(x_1, x_2) + d(x_2, x_3) + \ldots + d(x_{n-1}, x_n).$$

Exercise 5.4 Show that each of the following formulas defines a metric for $\mathbb{R}$:

$(a)\, d(x, y) = |x^3 - y^3|$, $(b)\, d(x, y) = |e^x - e^y|$, $(c)\, d(x, y) = |\tan^{-1}(x) - \tan^{-1}(y)|$.

Which property of the maps $x \mapsto x^3$, $x \mapsto e^x$, $x \mapsto \tan^{-1}(x)$ makes this work?

Exercise 5.5 Suppose that x, y are distinct points in a metric space (X, d) and let $\varepsilon = d(x, y)/2$. Prove that $B_\varepsilon(x)$ and $B_\varepsilon(y)$ are disjoint.

Exercise 5.6 Suppose that x, y are points in a metric space and that $\varepsilon > 0$. Show that if $y \in B_{\varepsilon/2}(x)$ then $B_{\varepsilon/2}(y) \subseteq B_\varepsilon(x)$.

Exercise 5.7 Show that if S is a bounded set in $\mathbb{R}^n$ then S is contained in $[a, b] \times [a, b] \times \ldots \times [a, b]$ for some $a, b \in \mathbb{R}$.

Exercise 5.8 Suppose that (X, d) is a metric space, $A \subseteq X$. Show that A is bounded iff there is some constant Δ such that $d(a, a') \leqslant \Delta$ for all $a, a' \in A$.

Exercise 5.9 Suppose that $A \subseteq B$ where B is a bounded subset of a metric space. Prove that A is bounded and diam $A \leqslant$ diam B

Exercise 5.10 Prove that if A, B are bounded subsets of a metric space and $A \cap B \neq \emptyset$ then diam $(A \cup B) \leqslant$ diam A + diam B.

Exercise 5.11 Sketch the open ball $B_1^{d_\infty}((0, 0))$ in $\mathbb{R}^2$.

Exercise 5.12 Suppose that d is a metric for a non-empty set X, and for any $x, y \in X$ define

$d^{(1)}(x, y) = kd(x, y)$, where k is a positive constant, $d^{(2)}(x, y) = \min\{1, d(x, y)\}$

$d^{(3)}(x, y) = d(x, y)/(1+d(x, y))$, $d^{(4)}(x, y) = d(x, y)^2$,

Prove that $d^{(1)}$, $d^{(2)}$, $d^{(3)}$ are metrics for X but $d^{(4)}$ may not be a metric for X.

Exercise 5.13 Prove that a subset of a metric space is open iff it is a union of open balls.

Exercise 5.14 Show that for any $x, y \in \mathbb{R}^n$,

$$d_\infty(x, y) \leqslant d_2(x, y) \leqslant d_1(x, y) \leqslant nd_\infty(x, y).$$

Exercise 5.15 Suppose that X is a non-empty set and that d, d' are metrics on X such that $d(x_1, x_2) \leqslant kd'(x_1, x_2)$ for all $x_1, x_2 \in X$ and some positive constant k.

(a) Show that $B_{\varepsilon/k}^{d'}(x) \subseteq B_\varepsilon^d(x)$ for any $x \in X$ and any $\varepsilon > 0$.

(b) Deduce that any subset of X which is d-open is also d'-open.

(c) Show that the open sets in $\mathbb{R}^n$ are the same for the metrics d_1, d_2, d_∞.

Exercise 5.16 Let (X, d_X) and (Y, d_Y) be metric spaces. As in Example 5.10, for $(x_1, y_1), (x_2, y_2) \in X \times Y$ let

$$d_1((x_1, y_1), (x_2, y_2)) = d_X(x_1, x_2) + d_Y(y_1, y_2),$$

$$d_2((x_1, y_1), (x_2, y_2)) = \sqrt{d_X(x_1, x_2)^2 + d_Y(y_1, y_2)^2}.$$

$$d_\infty((x_1, y_1), (x_2, y_2)) = \max\{d_X(x_1, x_2), d_Y(y_1, y_2)\}$$

(a) Prove that each of d_1, d_2, d_∞ is a metric for $X \times Y$

(b) Prove that for any $p, q \in X \times Y$,

$$d_\infty(p, q) \leqslant d_2(p, q) \leqslant d_1(p, q) \leqslant 2d_\infty(p, q).$$

(c) Show that the open sets in $X \times Y$ are the same for d_1, d_2, d_∞

(d) Let U, V be open subsets of X, Y respectively. Show that $U \times V$ is d_i-open in $X \times Y$ for $i = 1, 2, \infty$.

Exercise 5.17 Let (X, d) be a metric space and consider $X \times X$ as a metric space with the metric d_1 of Exercise 5.16. Show that $d : X \times X \to \mathbb{R}$ is continuous. (Hint: you could use Exercise 5.2.)

Exercise 5.18 Suppose that in a metric space X we have $B_r(x) = B_s(y)$ for some $x, y \in X$ and some positive real numbers r, s. Is $x = y$? Is $r = s$?

6 More concepts in metric spaces

We now explore some more concepts in metric spaces which generalize notions from real analysis and which in turn will be generalized to topological spaces later.

Closed sets

First we generalize the notion of a 'closed interval' $[a, b]$ in $\mathbb{R}$.

Definition 6.1 *A subset V of a metric space X is* closed in X *if $X \setminus V$ is open in X.*

Examples 6.2 (a) the following sets are all closed in $\mathbb{R}$:

(i) $[a, b]$, (ii) $(-\infty, 0]$, (iii) $\{0\}$, (iv) $\{1, 1/2, 1/3, \ldots, 1/n, \ldots\} \cup \{0\}$;

(b) the 'closed unit disc' $\{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 \leqslant 1\}$ is closed in $\mathbb{R}^2$;

(c) the 'closed rectangle' $\{(x_1, x_2) \in \mathbb{R}^2 : a \leqslant x_1 \leqslant b, c \leqslant x_2 \leqslant d\}$ is closed in $\mathbb{R}^2$.

(d) for a discrete metric space X, any subset of X is closed in X.

(e) in the space $\mathcal{C}([0, 1])$ of continuous real-valued functions on $[0, 1]$ with the sup metric, the subset $\{f \in \mathcal{C}([0, 1]) : f(1) = 0\}$ is closed. For its complement is open, since if $f(1) \neq 0$ then the same is true for all $g \in \mathcal{C}([0, 1])$ which are close enough to f in the sup metric e.g. such that $d_\infty(g, f) < |f(1)|/2$.

Intuitively, a set is closed if it contains all its 'boundary' points. We leave the proofs of Examples 6.2(a) (d) as exercises. The reader will quickly spot that results about closed sets can often be deduced from corresponding results about open sets by taking complements. This is true in particular for the next two results.

Proposition 6.3 *If $V_1, V_2, \ldots, V_m$ are closed subsets of a metric space X, then so is $\bigcup_{i=1}^{m} V_i$.*

Proposition 6.4 *The intersection of any family of sets each of which is closed in a metric space X is also closed in X.*

The proofs are a matter of taking complements and applying Proposition 5.39 and Proposition 5.41. Taking complements interchanges intersections and unions by De Morgan's Laws. For example, to prove Proposition 6.3 we observe that for each $i \in \{1, 2, \ldots, m\}$ the set $X \setminus V_i$ is open in X. Hence by Proposition 5.39 $\bigcap_{i=1}^{m}(X \setminus V_i)$ is open in X. But

$$X \setminus \bigcap_{i=1}^{m}(X \setminus V_i) = \bigcup_{i=1}^{m}(X \setminus (X \setminus V_i)) = \bigcup_{i=1}^{m} V_i, \quad \text{so} \quad \bigcup_{i=1}^{m} V_i \text{ is closed in } X.$$

Likewise from Example 5.34 we easily deduce

Proposition 6.5 *For any metric space X, the empty set $\emptyset$ and the whole set X are closed in X.*

We should warn that subsets of a metric space X are 'nothing like doors'—that is to say, many subsets of X are neither open nor closed in X, and we can have a subset which is both open and closed in X; for examples, think of $[0, 1)$ in $\mathbb{R}$ and the whole set X in any metric space X.

We can express continuity in terms of closed sets as well as in terms of open sets.

Proposition 6.6 *Let X, Y be metric spaces and let $f : X \to Y$ be a map. Then f is continuous iff $f^{-1}(V)$ is closed in X whenever V is closed in Y.*

Proof The proof follows from Proposition 5.37 by taking complements. (Remember that $X \setminus f^{-1}(V) = f^{-1}(Y \setminus V)$.) □

Closure

We have already remarked that a general subset of a metric space X is likely to be neither open nor closed in X. However, we can get from an arbitrary subset of X to one that is closed in X and also to one that is open in X, in rather natural ways. We shall explain first how to get from a general subset A of X to a related set, called the closure of A in X, which is closed in X. Intuitively, to get from A to its closure, written $\overline{A}$, we add in all points of X which are 'arbitrarily close to A'.

Definition 6.7 *Suppose that A is a subset of a metric space X, and $x \in X$. We say that x* is a point of closure of A in X *if given $\varepsilon > 0$*

we have $B_\varepsilon(x) \cap A \neq \emptyset$. *The* closure of A in X *is the set of all points of closure of* A *in* X.

When it is agreed which metric space X we are taking closures in, we denote the closure of A in X by $\overline{A}$.

Example 6.8 (a) The closure of each of the intervals $(0, 1)$, $[0, 1)$, $(0, 1]$, $[0, 1]$ in $\mathbb{R}$ is the interval $[0, 1]$.

(b) The closure of $B_1((0, 0))$ in $\mathbb{R}^2$ is $\{x \in \mathbb{R}^2 : d_2(x, 0) \leqslant 1\}$.

(c) If A is a non-empty bounded subset of $\mathbb{R}$ then $\sup A$ and $\inf A$ are in $\overline{A}$.

Closure gives rise to a concept which is often important in analysis.

Definition 6.9 *A subset* A *of a metric space* X *is said to be* dense in X *if* $\overline{A} = X$.

Example 6.10 Both the set $\mathbb{Q}$ of rational numbers and the set $\mathbb{R} \setminus \mathbb{Q}$ of irrational numbers are dense in $\mathbb{R}$.

The next proposition is a survey of properties of closure.

Proposition 6.11 *Let* A, B *be subsets of a metric space* X. *Then*

(*a*) $A \subseteq \overline{A}$;

(*b*) $A \subseteq B$ *implies that* $\overline{A} \subseteq \overline{B}$;

(*c*) A *is closed in* X *if and only if* $\overline{A} = A$;

(*d*) $\overline{\overline{A}} = \overline{A}$;

(*e*) $\overline{A}$ *is closed in* X;

(*f*) $\overline{A}$ *is the smallest closed subset of* X *containing* A.

Proof Properties (a) and (b) are clear after a little thought. To prove (c), suppose first that A is closed in X. We shall show that no point of its complement is in $\overline{A}$. For $X \setminus A$ is open in X, so if $x \in X \setminus A$ then there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq X \setminus A$, so $B_\varepsilon(x) \cap A = \emptyset$. This shows that $x \notin \overline{A}$. Thus $\overline{A} \subseteq A$, and since $A \subseteq \overline{A}$ by (a), we get $\overline{A} = A$.

Conversely if $\overline{A} = A$ we can show that $X \setminus A$ is open in X, hence A is closed in X. For if $x \in X \setminus A$ then $x \notin \overline{A}$, so for some $\varepsilon > 0$ we have $B_\varepsilon(x) \cap A = \emptyset$, so $B_\varepsilon(x) \subseteq X \setminus A$, and the latter is open as required.

To prove (d), first note that $\overline{A} \subseteq \overline{\overline{A}}$ by (b). Now let $x \in \overline{\overline{A}}$ and let $\varepsilon > 0$. Then $B_{\varepsilon/2}(x) \cap \overline{A} \neq \emptyset$, so there is some $y \in B_{\varepsilon/2}(x) \cap \overline{A}$. Now $B_{\varepsilon/2}(y) \subseteq B_\varepsilon(x)$ (see Exercise 5.6). Also $B_{\varepsilon/2}(y) \cap A \neq \emptyset$ since $y \in \overline{A}$. Hence $B_\varepsilon(x) \cap A \neq \emptyset$, so $x \in \overline{A}$. This shows that $\overline{\overline{A}} \subseteq \overline{A}$, so $\overline{\overline{A}} = \overline{A}$.

Now (e) follows from (c) and (d).

Finally we expand on what (f) means and then prove it. It means that $A \subseteq \overline{A}$ (which we know already from (a)), and also that if B is any closed subset of X satisfying $A \subseteq B$ then $\overline{A} \subseteq B$. So suppose B is such a closed set. By (b) we have $\overline{A} \subseteq \overline{B}$. But by (c) we have $\overline{B} = B$. So $\overline{A} \subseteq B$ as required. □

Another way of expressing (f) is to say that $\overline{A}$ is the intersection, call it V, of all sets which are both closed in X and contain A. For, by Proposition 5.39, V is also closed in X, and V contains A, so $\overline{A} \subseteq V$. But $\overline{A}$ is itself closed in X by (c), and contains A by (a), so it is one of the sets we take the intersection of to form V, hence $V \subseteq \overline{A}$. This proves that $V = \overline{A}$.

We have earlier expressed continuity of a map $f : X \to Y$ of metric spaces in terms of open sets and in terms of closed sets. We can also express it in terms of closure.

Proposition 6.12 *A map $f : X \to Y$ of metric spaces is continuous if and only if $f(\overline{A}) \subseteq \overline{f(A)}$ for every $A \subseteq X$.*

Proof This is an exercise at the end of the chapter. □

There are analogues of Propositions 6.3, 6.4 for closure.

Proposition 6.13 *Let $A_1, A_2, \ldots, A_m$ be subsets of a metric space X. Then*

$$\overline{\bigcup_{i=1}^{m} A_i} = \bigcup_{i=1}^{m} \overline{A_i}.$$

Proposition 6.14 *For each i in some indexing set I, let A_i be a subset of the metric space X. Then*

$$\overline{\bigcap_{i \in I} A_i} \subseteq \bigcap_{i \in I} \overline{A_i}.$$

Proof The proofs are exercises at the end of the chapter. □

Equality does not necessarily hold in Proposition 6.14 even when the index set is finite (see Exercise 6.15).

Limit points

For analysis it is often useful to consider limit points as well as closure.

Definition 6.15 *A point x in a metric space X is said to be* a limit point *of a subset A in X if given $\varepsilon > 0$ there is a point in $B_\varepsilon(x) \cap A$ other than x itself, i.e. $(B_\varepsilon(x) \setminus \{x\}) \cap A \neq \emptyset$.*

Another name sometimes used for this is: x is a point of accumulation of A in X. Notice the difference here from the definition of a point of closure: it is not enough for x itself to be in A. Thus points in A may or may not be limit points of A although they are always points of closure of A. On the other hand it follows immediately from the definition that all limit points of A in X are in $\overline{A}$.

Example 6.16 (a) Let $A = [0, 1) \cup \{2\}$. Then the limit points of A in $\mathbb{R}$ are the points of $[0, 1]$, while $\overline{A} = [0, 1] \cup \{2\}$.

(b) Let $A = \{1/n : n \in \mathbb{N}\} \cup \{0\}$. Then A has only one limit point in $\mathbb{R}$, namely 0, while $\overline{A} = A$.

The main use of limit points is to recognize which sets in a metric space X are closed in X, using the next proposition.

Proposition 6.17 *A subset A of a metric space X is closed in X iff it contains all its limit points in X.*

Proof This is a corollary of the next proposition. □

Proposition 6.18 *Let A be any subset of a metric space X. Then $\overline{A}$ is the union of A with all its limit points in X.*

Proof Let us temporarily write B for the union of A with all its limit points in X. We have already noted that from the definitions all limit points of A in X are contained in $\overline{A}$. By Proposition 6.11 $A \subseteq \overline{A}$. Thus $B \subseteq \overline{A}$.

Conversely suppose x is in $\overline{A}$. If $x \in A$ then $x \in B$ as required. Suppose $x \notin A$. Then since $x \in \overline{A}$, for any $\varepsilon > 0$ we know $B_\varepsilon(x) \cap A \neq \emptyset$, and since $x \notin A$ this tells us that $(B_\varepsilon(x) \setminus \{x\}) \cap A \neq \emptyset$, so x is a limit point of A in X, and again $x \in B$. This proves $\overline{A} \subseteq B$. So $B = \overline{A}$ as required. □

It is not usually a good idea to use this result in proving facts about closures if we set out with the 'definition' of closure of A in X as the union of A with its limit points in X, then we tend to have to distinguish two cases in our arguments, whereas Definition 6.7 usually allows a more streamlined treatment.

Interior

The idea of interior is, in a sense which we shall try to explain, dual to the idea of closure. The closure of a subset A adds in all points which are

intuitively very close to A, whereas the interior of A consists of all points which are 'well inside' A.

Definition 6.19 *The* interior $\mathring{A}$ *of a subset A in a metric space X is the set of points $a \in A$ such that $B_\varepsilon(a) \subseteq A$ for some $\varepsilon > 0$.*

Example 6.20 (a) the interior of any of the intervals (a, b), $[a, b)$, $(a, b]$, $[a, b]$ in $\mathbb{R}$ is (a, b).

(b) the interior of $\mathbb{Q}$ in $\mathbb{R}$ is $\emptyset$.

Comparison of the next proposition with Proposition 6.11—especially their parts (f)—indicates the sense in which interior is dual to closure.

Proposition 6.21 *Let A, B be subsets of a metric space X. Then*

(a) $\mathring{A} \subseteq A$;

(b) $A \subseteq B$ *implies that* $\mathring{A} \subseteq \mathring{B}$;

(c) A *is open in* X *iff* $\mathring{A} = A$;

(d) $\mathring{\mathring{A}} = \mathring{A}$;

(e) $\mathring{A}$ *is open in* X;

(f) $\mathring{A}$ *is the largest open subset of* X *contained in* A.

Proof (a) If $x \in \mathring{A}$ then by definition $x \in A$.

(b) If $A \subseteq B$ and $x \in \mathring{A}$ then there exists $\varepsilon > 0$ with $B_\varepsilon(x) \subseteq A \subseteq B$ so $x \in \mathring{B}$.

(c) If A is open in X then it follows from the definition that there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq A$ so $x \in \mathring{A}$. This shows that then $A \subseteq \mathring{A}$, and since $\mathring{A} \subseteq A$ by (a), we get $\mathring{A} = A$. Conversely if $\mathring{A} = A$ then any $x \in A$ is in $\mathring{A}$, so there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq A$; this shows that A is open in X.

(d) From (b), $\mathring{\mathring{A}} \subseteq \mathring{A}$. Conversely suppose $x \in \mathring{A}$. Then there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq A$. Now for any $y \in B_\varepsilon(x)$ we know (cf. Proposition 5.31) there exists $\delta > 0$ such that $B_\delta(y) \subseteq B_\varepsilon(x) \subseteq A$, hence $y \in \mathring{A}$. This shows that

$B_\varepsilon(x) \subseteq \mathring{A}$, so $x \in \mathring{\mathring{A}}$. Hence $\mathring{A} \subseteq \mathring{\mathring{A}}$, and $\mathring{\mathring{A}} = \mathring{A}$ as required.

(e) What this means is that $\mathring{A}$ is open, which follows from (c) and (d), and also that if $B \subseteq A$ is open in X then $B \subseteq \mathring{A}$. This follows since B is open in X so $\mathring{B} = B$, and $\mathring{B} \subseteq \mathring{A}$ from (b). □

Remark We could also prove Proposition 6.21 by relating interior to closure and deducing Proposition 6.21 from Proposition 6.11. For variety, we shall follow that route when we return to these ideas in the more general context of Chapter 7.

As Exercise 6.20 asserts, continuity in metric spaces may also be characterised in terms of interior.

Boundary

Next we consider the *boundary* or *frontier* of a subset A in a metric space X.

Definition 6.22 *The* boundary ∂A *of a subset A in a metric space X is the set* $\overline{A} \setminus \mathring{A}$.

Intuitively, the boundary consists of all points which are 'very close' to A but are not in the interior of A. Alternatively, in view of Proposition 6.24 below, we may think of ∂A as consisting of points which are very close to both A and $X \setminus A$.

In the next example, (a) illustrates the standard situation, in line with our intuition, whereas (b) shows that the definition can produce some surprising examples.

Example 6.23 (a) In $\mathbb{R}$, the boundary of each interval (a, b), $[a, b)$, $(a, b]$, $[a, b]$ is $\{a, b\}$. The boundary of $[0, 1] \cup \{2\}$ is $\{0, 1, 2\}$.

(b) The boundary of $\mathbb{Q}$ in $\mathbb{R}$ is $\mathbb{R}$. (This is Exercise 6.21).

We shall not prove much about boundaries here, but the next proposition provides an alternative definition of boundary.

Proposition 6.24 *Given a subset A of a metric space X, a point $x \in X$ is in ∂A iff for every $\varepsilon > 0$ both $A \cap B_\varepsilon(x)$ and $(X \setminus A) \cap B_\varepsilon(x)$ are non-empty.*

Proof Suppose $x \in \partial A$ and let $\varepsilon > 0$. Since $x \in \overline{A}$ it follows by definition of $\overline{A}$ that $A \cap B_\varepsilon(x) \neq \emptyset$. But also $x \notin \mathring{A}$ so $B_\varepsilon(x) \not\subseteq A$, and this shows that $(X \setminus A) \cap B_\varepsilon(x) \neq \emptyset$.

Conversely suppose that both $A \cap B_\varepsilon(x)$ and $(X \setminus A) \cap B_\varepsilon(x)$ are non-empty, for any choice of $\varepsilon > 0$. From the first of these we get that $x \in \overline{A}$ and from the second, that $x \notin \mathring{A}$. □

A few more intuitively plausible properties of boundary are left as Exercise 6.23.

Convergence in metric spaces

As in Chapter 4 a sequence of objects in any set X may be defined formally as a map $s : \mathbb{N} \to X$, but we use the traditional notation (x_n) for a sequence, putting $x_n = s(n)$.

Definition 6.25 *A sequence (x_n) in a metric space X* converges *to a point $x \in X$ if given (any real number) $\varepsilon > 0$, there exists (an integer) N such that $x_n \in B_\varepsilon(x)$ whenever $n \geqslant N$.*

Just as for real number sequences, we get uniqueness of limits for sequences in metric spaces.

Proposition 6.26 *Suppose that a sequence (x_n) in a metric space (X, d) converges to x and also to y in X. Then $x = y$.*

Proof This is just like the proof of Proposition 4.13. Suppose that the sequence (x_n) in X converges both to x and to y. Suppose $y \neq x$, and let $\varepsilon = d(x, y)/2$. Then $B_\varepsilon(x)$ and $B_\varepsilon(y)$ are disjoint, by Exercise 5.5, and since x_n is supposed to belong to each of these for sufficiently large n, we get a contradiction. □

Just as for real sequences we can also consider Cauchy sequences in a metric space.

Definition 6.27 *A sequence (x_n) in a metric space (X, d) is called* a Cauchy sequence *if for $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $d(x_m, x_n) < \varepsilon$ whenever $m, n \geqslant N$ (recall this means $m \geqslant N$ and $n \geqslant N$).*

The proof of the following result (Exercise 6.24) is entirely similar to that of the easy part of Theorem 4.18. Its converse will be considered in Chapter 17.

Proposition 6.28 *Any convergent sequence in a metric space is a Cauchy sequence.*

We prove just one more result about convergence in metric spaces in the meantime.

Proposition 6.29 *Suppose that Y is a subset of a metric space X and that (y_n) is a sequence in Y which converges to a point $a \in X$. Then $a \in \overline{Y}$.*

Proof Since (y_n) converges to a, for any $\varepsilon > 0$ we have $y_n \in B_\varepsilon(a)$ for all sufficently large n, and y_n is a point of Y. Hence a is in the closure $\overline{Y}$ of Y in X. □

Corollary 6.30 *If Y is a closed subset of a metric space X and (y_n) is a sequence of points in Y which converges in X to a point a then $a \in Y$.*

Equivalent metrics

We introduced metrics to study continuity, so it is reasonable to call two metrics on a set equivalent if they make the same maps in to and out of that set continuous. The next definition achieves that, as we prove in Proposition 6.32. Although the word *topology* will not be defined until the next chapter, the name *topologically equivalent* will be used here to avoid confusion with another equivalence relation on metrics.

Definition 6.31 *Metrics d_1 and d_2 on a set X are called* topologically equivalent *if a subset U of X is d_1-open in X iff it is d_2-open in X.*

Paraphrasing this definition, metrics are topologically equivalent if they make the same sets open. Topological equivalence of metrics is clearly an equivalence relation.

Proposition 6.32 *Suppose that d_1, d_2 are topologically equivalent metrics for a set X. For any metric spaces (Y, d_Y), (Z, d_Z) and for any maps $f : Y \to X$, $g : X \to Z$ the following hold:*

(a) f is (d_Y, d_1)-continuous iff it is (d_Y, d_2)-continuous;

(b) g is (d_1, d_Z)-continuous iff it is (d_2, d_Z)-continuous.

Proof Suppose that $f : Y \to X$ is (d_Y, d_1)-continuous. Then, by Proposition 5.37, $f^{-1}(U)$ is d_Y-open in Y whenever U is a d_1-open subset of X. Now let U be a d_2-open subset of X; by topological equivalence U is also d_1-open, so $f^{-1}(U)$ is d_Y-open in Y. Again by Proposition 5.37 this shows that f is (d_Y, d_2)-continuous. The converse, that if f is (d_Y, d_2)-continuous then it is (d_Y, d_1)-continuous, is proved similarly.

The part concerning g is also proved similarly: suppose that $g : X \to Z$ is (d_1, d_Z)-continuous. Then, by Proposition 5.37, $g^{-1}(U)$ is a d_1-open subset of X whenever U is a d_Z-open subset of Z. By topological equivalence of d_1 and d_2 it follows that $g^{-1}(U)$ is a d_2-open subset of X whenever U is a d_Z-open subset of Z. Hence, again by Proposition 5.37, $g : X \to Z$ is (d_2, d_Z)-continuous. The proof of the converse, that if $g : X \to Z$ is (d_2, d_Z)-continuous then it is (d_1, d_Z)-continuous, is entirely similar. □

Quite often when two metrics are topologically equivalent it is because they are equivalent in a stronger way, which we define next.

Definition 6.33 *Two metrics d_1, d_2 on a set X will be called* Lipschitz equivalent *if there are positive constants h, k such that for any x, $y \in X$,*

$$hd_2(x, y) \leqslant d_1(x, y) \leqslant kd_2(x, y).$$

The reader should be warned that although the name we have chosen for the concept seems appropriate, it is not universally used. It is clear that Definition 6.33 gives an equivalence relation.

Proposition 6.34 *Lipschitz equivalent metrics are topologically equivalent.*

Proof Suppose that metrics d_1, d_2 on X satisfy Definition 6.33. Then for any $\varepsilon > 0$ and any $x \in X$ the following inclusions hold (see Exercise 5.15).

$$B^{d_1}_{h\varepsilon}(x) \subseteq B^{d_2}_{\varepsilon}(x), \quad B^{d_2}_{\varepsilon/k}(x) \subseteq B^{d_1}_{\varepsilon}(x).$$

(Warning: it is easy to get these the wrong way round!) Now suppose that $U \subseteq X$ is d_1-open, and let $x \in U$. Then there is an $\varepsilon > 0$ such that $B^{d_1}_{\varepsilon}(x) \subseteq U$. Hence $B^{d_2}_{\varepsilon/k}(x) \subseteq B^{d_1}_{\varepsilon}(x) \subseteq U$. This proves that U is d_2-open. The converse is proved similarly. □

Example 6.35 Let $X = \mathbb{R}^n$ and let d_1, d_2, d_∞ be as in Example 5.7. Then (see Exercise 5.14) for all x, $y \in \mathbb{R}^n$ we have

$$d_\infty(x, y) \leqslant d_2(x, y) \leqslant d_1(x, y) \leqslant nd_\infty(x, y).$$

This shows that all three of these metrics are Lipschitz equivalent. Hence by Proposition 6.34 they are topologically equivalent. Thus although the open balls with respect to these three metrics differ in shape, the open sets are the same. In a particular situation, it may be more convenient to use one of these metrics rather than the others. We now know that different choices will not affect anything that concerns continuity alone. Likewise the metrics in Example 5.10 are Lipschitz and hence topologically equivalent. This finally justifies the claims made in the previous chapter that propositions about products proved using d_1 are true also using d_2 or d_∞.

Not all metrics on $\mathbb{R}^n$ are equivalent, however; we have already seen that the discrete metric makes singleton sets open, so it cannot be topologically equivalent to d_1, d_2, d_∞. More interesting examples of non-equivalent metrics arise in function spaces.

Example 6.36 On the set $C[a, b]$ of all continuous real-valued functions on $[a, b]$ the sup metric d_∞ and the L^1 metric d_1 are not topologically

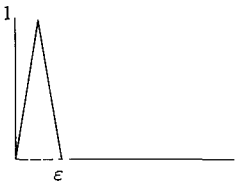

Figure 6.1. Graph of f

equivalent. To see this, let $f, g \in C[a, b]$. Then $|f(t) - g(t)| \leqslant d_\infty(f, g)$ for all $t \in [a, b]$ by definition of d_∞. Hence by integration theory,

$$d_1(f, g) = \int_a^b |f(t) - g(t)| \, \mathrm{d}t \leqslant (b - a) d_\infty(f, g),$$

which is 'half of a Lipschitz equivalence' and gives $B_\varepsilon^{d_\infty}(f) \subseteq B_{(b-a)\varepsilon}^{d_1}(f)$ for any $f \in C[a, b]$ and any $\varepsilon > 0$.

However, if we let 0 denote also the constant function with value 0, then $B_1^{d_\infty}(0)$, which is d_∞-open, is not d_1-open. For if it were then we would have $B_\varepsilon^{d_1}(0) \subseteq B_1^{d_\infty}(0)$ for some $\varepsilon > 0$. But for any $\varepsilon > 0$ there exists a continuous function f on $[a, b]$ such that $d_1(f, 0) < \varepsilon$ yet $f \notin B_1^{d_\infty}(0)$ (see Figure 6.1 for the graph of such a function).

Closely related to these definitions of equivalence of metrics we have also equivalences between metric spaces, say (X, d_X) and (Y, d_Y).

Definition 6.37 *A* topological equivalence *or* homeomorphism *is a bijective map $f : X \to Y$ such that f and f^{-1} are both continuous.*

Definition 6.38 *A* Lipschitz equivalence *is a bijective map $f : X \to Y$ such that there exist strictly positive constants h, k satisfying*

$$h d_Y(f(x_1), f(x_2)) \leqslant d_X(x_1, x_2) \leqslant k d_Y(f(x_1), f(x_2))$$

for all $x_1, x_2 \in X$.

Again, we note that the terminology *Lipschitz* is not universally used for this concept.

As before, a Lipschitz equivalence is a topological equivalence. A special case of a Lipschitz equivalence between metric spaces (X, d_X) (Y, d_Y) is given in the next definition.

Definition 6.39 *An* isometry $f : X \to Y$ *is a bijective map such that*

$$d_Y(f(x_1), f(x_2)) = d_X(x_1, x_2) \text{ for all } x_1, x_2 \in X.$$

We also use the term *isometry into* for a map which satisfies the conditions for an isometry except that it is not necessarily onto.

Next we give a familiar example of an isometry.

Example 6.40 Let $f : \mathbb{R}^2 \to \mathbb{C}$ be given by $f(x_1, x_2) = x_1 + ix_2$. Then f is an isometry from $\mathbb{R}^2$ with the Euclidean metric to $\mathbb{C}$ with the metric of Example 5.5.

The precise relation between Definitions 6.31, 6.33 on the one hand and Definitions 6.37, 6.38 on the other is the following: two metrics d_1, d_2 on a set X are topologically (resp. Lipschitz) equivalent if and only if the identity map of X is a homeomorphism (resp. Lipschitz equivalence) from (X, d_1) to (X, d_2). Likewise, if f is a homeomorphism (resp. Lipschitz equivalence) from a metric space (X, d_X) to (Y, d_Y) then the formula $d(x_1, x_2) = d_Y(f(x_1), f(x_2))$ defines a metric d on X which is topologically (resp. Lipschitz) equivalent to d_X.

Review

In this chapter and the previous one we began by looking at metric spaces as a general framework in which to study continuity. However, we have seen in Proposition 5.37 and Proposition 6.32 that it is not so much the particular metrics on spaces that determine which maps between them are continuous, but the topological equivalence classes of these metrics, in other words, the 'open set structure' defined by the metrics. So in order to define continuity of a map from one set to another, perhaps after all we do not really need metrics on the sets, but only some adequate notion of 'open subset'. This leads us to topological spaces, the subject of the next chapter. In it we shall follow the same pattern as the move from Euclidean spaces to general metric spaces. There we first observed that continuity of functions between Euclidean spaces can be expressed in terms of Euclidean distance. We then wrote down three properties of Euclidean distance and used them as *axioms* for metric spaces. Likewise now we write down some properties of open sets in metric spaces, and use them as *axioms* for topological spaces. The properties of open sets in metric spaces that we choose is again a matter of historical trial and error, as in the case of the metric space axioms. Again we shall go straight to the historical winners in the next chapter.

There is another important property, completeness, which we shall study in the context of metric spaces; we return to that in Chapter 16.

Exercise 6.1 Check that the sets in Example 6.2 are closed in $\mathbb{R}$.

Exercise 6.2 Which of the following sets are closed in $\mathbb{R}$?
(a) $[1, \infty)$, (b) $\mathbb{R} \setminus \mathbb{Q}$, (c) $\{n/(n+1) : n \in \mathbb{N}\}$,
(d) $\{1/n : n \in \mathbb{N},\ n \geqslant 2\} \cup \{0, 1, 2\}$.

Exercise 6.3 Prove that any finite subset of a metric space X is closed in X.

Exercise 6.4 Prove Proposition 6.4, that the intersection of any family of sets each closed in a metric space X is also closed in X.

Exercise 6.5 Let $C_0 = [0, 1]$, let C_1 denote C_0 with its open middle third removed, so that C_1 is the union of the two closed intervals $[0, 1/3]$, $[2/3, 1]$ each of length $1/3$. Inductively suppose we have defined C_n as the union of 2^n closed intervals each of length $1/3^n$, and define C_{n+1} to be the result of removing from each interval in C_n its open middle third, so that C_{n+1} is the union of 2^{n+1} closed intervals each of length $1/3^{n+1}$. Finally let $C = \bigcap_{n=0}^{\infty} C_n$. Show that C (which is called 'the Cantor middle-third set') is closed in $\mathbb{R}$.

Exercise 6.6 Let $\mathcal{C}[0, 1]$ be the space of continuous real-valued functions on $[0, 1]$ with the sup metric, and $A \subseteq [0, 1]$. Show that the following subset is closed in $\mathcal{C}[0, 1]$: $\{f \in \mathcal{C}[0, 1] : f(a) = 0 \text{ for all } a \in A\}$.

Exercise 6.7 Prove that the closure of each of the intervals $(0, 1)$, $[0, 1)$, $(0, 1]$, and $[0, 1]$ in $\mathbb{R}$ is $[0, 1]$.

Exercise 6.8 Prove that, in $\mathbb{R}^2$, $\overline{B_1((0, 0))} = \{x \in \mathbb{R}^2 : d_2(x, 0) \leqslant 1\}$.

Exercise 6.9 Prove that if A is a non-empty bounded subset of $\mathbb{R}$ then $\sup A$ and $\inf A$ are in $\overline{A}$.

Exercise 6.10 Prove that if A is a bounded subset of a metric space then $\overline{A}$ is bounded and $\operatorname{diam} \overline{A} = \operatorname{diam} A$.

Exercise 6.11 Describe the closure of each of the sets in Exercise 6.2.

Exercise 6.12 Let x be a point in a metric space (X, d) and let $r > 0$ be a real number. Define $\overline{B}_r(x) = \{y \in X : d(y, x) \leqslant r\}$. Show that $\overline{B}_r(x)$ is closed in X. Show also that $\overline{B_r(x)} \subseteq \overline{B}_r(x)$ and give an example to show that the inclusion may be strict.
[Hint: Think about discrete metric spaces.]

Exercise 6.13 Prove Proposition 6.12, that a map $f : X \to Y$ of metric spaces is continuous iff $f(\overline{A}) \subseteq \overline{f(A)}$ for all subsets $A \subseteq X$.

Exercise 6.14 Prove Proposition 6.13, that if $A_1, A_2, \ldots, A_m$ are subsets of a metric space X, then

$$\overline{\bigcup_{i=1}^{m} A_i} = \bigcup_{i=1}^{m} \overline{A_i}.$$

Exercise 6.15 Prove Proposition 6.14, that if for each i in some indexing set I we have a subset A_i of a metric space X, then

$$\overline{\bigcap_{i \in I} A_i} \subseteq \bigcap_{i \in I} \overline{A_i}.$$

Give an example to show that equality may fail even when there are only two sets involved.

Exercise 6.16 For a point x and a non-empty subset A of a metric space (X, d), define $d(x, A) = \inf\{d(x, a) : a \in A\}$.

(a) Prove that $d(x, A) = 0$ iff $x \in \overline{A}$.

(b) Show that if y is another point in X then $d(x, A) \leqslant d(x, y) + d(y, A)$.

(c) Prove that $x \mapsto d(x, A)$ gives a continuous map from X to $\mathbb{R}$.

Exercise 6.17 Describe the set of limit points in $\mathbb{R}$ of each of the sets in Exercise 6.2.

Exercise 6.18 Prove that a finite subset of a metric space has no limit points.

Exercise 6.19 Prove that Proposition 6 17 follows from Proposition 6.18

Exercise 6.20 Prove that a map $f : X \to Y$ of metric spaces is continuous iff for every subset $B \subseteq Y$ we have that $f^{-1}\left(\overset{\circ}{B}\right)$ is contained in the interior of $f^{-1}(B)$.

Exercise 6.21 Prove that

(a) the boundary of each interval (a, b). $[a, b)$. $(a. b]$, $[a, b]$ is $\{a, b\}$,

(b) the boundary of $\mathbb{Q}$ in $\mathbb{R}$ is $\mathbb{R}$.

Exercise 6.22 Given a non-empty subset A of a metric space X, prove that a point $x \in X$ is in ∂A iff $d(x, A) = 0 = d(x, X \setminus A)$, where $d(x, A)$ is defined as in Exercise 6.16.

Exercise 6.23 For a subset A of a metric space X, prove:

(a) $\mathring{A} = A \setminus \partial A = \overline{A} \setminus \partial A$,
(b) $\overline{X \setminus A} = X \setminus \mathring{A}$,
(c) $\partial A = \overline{A} \cap \overline{X \setminus A} = \partial(X \setminus A)$,
(d) ∂A is closed in X.

Exercise 6.24 Prove that any convergent sequence in a metric space is Cauchy.

Exercise 6.25 Prove that the following is a characterization of continuity for a map $f : X \to Y$ of metric spaces: whenever (x_n) is a sequence in X converging to a point $x \in X$ we have that $(f(x_n))$ converges to $f(x)$ in Y.

Exercise 6.26 Prove the following converse to Proposition 6.29: Suppose that Y is a subset of a metric space X and that $a \in \overline{Y}$. Prove that there is a sequence in Y which converges in X to a. Deduce that if every sequence in Y which converges in X has its limit in Y then Y is closed in X.

Exercise 6.27 Prove that the metrics $d^{(2)}$, $d^{(3)}$ in Exercise 5.12 are topologically equivalent to d.

7 Topological spaces

In this chapter, we make our final leap into generality: we introduce topological spaces as our ultimate framework for studying continuity. At the end of the last chapter we saw that the open sets in a metric space are the most important elements when defining continuity. In the light of the remarks there, the following is a plausible definition.

Definition

Definition 7.1 *A* topological space $T = (X, \mathcal{T})$ *consists of a non-empty set* X *together with a fixed family* $\mathcal{T}$ *of subsets of* X *satisfying*

(*T1*) $X, \emptyset \in \mathcal{T}$,

(*T2*) *the intersection of any two sets in* $\mathcal{T}$ *is in* $\mathcal{T}$,

(*T3*) *the union of any collection of sets in* $\mathcal{T}$ *is in* $\mathcal{T}$.

The family $\mathcal{T}$ is called a *topology for* X, and the members of $\mathcal{T}$ are called the *open sets of* T. Thus '$U \in \mathcal{T}$' and 'U is open in T' mean the same thing. Elements of X are called *points in the space* T. In practice we often use the same name for X and T when it is unmistakable which topology is intended. Thus in the above context we refer to 'the topological space X', 'points of X', and 'open sets of X'. In fact, as we go on we shall refer to just the 'space' X when it is clear that this is short for 'topological space'. All this is intended to avoid some clumsy notation. However, there is always a topology $\mathcal{T}$ understood, and whenever it is desirable for clarity we shall use the full notation $(X, \mathcal{T})$.

From (T2) it follows by induction that the intersection of any finite collection of open sets in X is open in X. Note that one cannot get to a statement about an *infinite* intersection (or union) by induction from statements about finite intersections (or unions). In general, as we shall see, an infinite intersection of open sets in a topological space is not open.

It is important to remember that $\mathcal{T}$ is in general only a *sub*family of the family of *all* subsets of X.

We note here an easy result which we often use while proving that some set is open.

Proposition 7.2 *For a subset U of a topological space X to be open in X it is necessary and sufficient that for every $x \in U$ there is an open subset U_x of X such that $x \in U_x \subseteq U$.*

Proof If U is open in X, then for each $x \in U$ we may take $U_x = U$ and the condition holds.

Conversely suppose the condition holds. We shall check that

$$U = \bigcup_{x \in U} U_x. \qquad (**)$$

Thus U is a union of sets open in X so U is open in X. To prove $(**)$ we first suppose $x_0 \in U$. Then $x_0 \in U_{x_0} \subseteq \bigcup_{x \in U} U_x$. Conversely, any point in the union is in U_x for some $x \in U$, and we know $U_x \subseteq U$ so $x \in U_x \subseteq U$. □

Note that we have insisted that the set X in a topological space $(X, \mathcal{T})$ should be non-empty. The reason for this choice is the same as for metric spaces – it saves us from having to add the condition of being non-empty as a requirement in certain results later on.

Examples

We now give several examples of topological spaces. The first is the class of examples that motivated the definition of topological spaces.

Example 7.3 Any metric space (X, d) gives rise to a topological space $(X, \mathcal{T}_d)$ where $\mathcal{T}_d$ is defined to be the family of all d-open subsets of X, i.e. those subsets of X which are open in X according to Definition 5.32. By Example 5.34, Proposition 5.39, and Proposition 5.41, the topological space axioms are satisfied by this family $\mathcal{T}_d$.

This use of the term *open* in two slightly different contexts is confusing at first. On the one hand, when a metric space (X, d) has been specified, we can work out, using Definition 5.32, whether or not a given subset of X is d-open. On the other hand, when a topological space $(X, \mathcal{T})$ has been specified, then to tell whether a given subset U of X is open in X (i.e. whether $U \in \mathcal{T}$), we have only to 'look' at the list $\mathcal{T}$ and check whether U is on it. A graphic comparison may help fix this distinction: a nightclub bouncer may have a list of criteria (wearing a blouse or shirt, not wearing jeans or trainers, etc.) to work through before deciding whether to admit you. But a doorkeeper at a private party may just have a list of those to be admitted: if you're on the list, you're in. The bouncer's decision is like

deciding whether a subset of a given metric space X is open in X, while the doorkeeper's nod is like deciding that a subset of a given topological space X is open in X.

Example 7.3 says that given a metric space (X, d) we may construct a topological space $(X, \mathcal{T}_d)$ by defining $\mathcal{T}_d$ to consist of precisely those subsets of X which are d-open. A topological space which arises in this way from a metric space is called *metrizable*. In this case, the two meanings of open coincide (being d-open in the metric space (X, d) and being open in the topological space $(X, \mathcal{T}_d)$). We call $(X, \mathcal{T}_d)$ the topological space *underlying* the metric space (X, d), and we call $\mathcal{T}_d$ the topology *induced by* the metric d.

Our discussion of topologically equivalent metrics in the previous chapter shows that distinct metric spaces may give rise to the same topological space. For example, the metrics d_1, d_2, d_∞ on $\mathbb{R}^n$ all give rise to the same open sets and hence to the same topology ('the Euclidean topology'). Whenever we refer to $\mathbb{R}^n$ or a subspace of $\mathbb{R}^n$ as a topological space, this topology will be understood unless some other is specified.

After the warning above that $\mathcal{T}$ is in general only a *sub*family of the family of subsets of X, the next example may seem perverse. This example plays a similar role to the discrete metric in providing counterexamples.

Example 7.4 Let X be any non-empty set, and let $\mathcal{T}$ be the set of *all* subsets of X. The topological space axioms are clearly satisfied by this $\mathcal{T}$. We call $\mathcal{T}$ the *discrete topology* on X and the resulting space $(X, \mathcal{T})$ a *discrete* space.

Example 7.4 is actually a special case of Example 7.3, for the discrete topology on X is induced by the discrete metric on X –this follows from Example 5.35.

Example 7.4 leads us to think about the other extreme form of topology.

Example 7.5 Let X be any non-empty set. The *indiscrete topology* on X is the family $\{\emptyset, X\}$.

Proof It is easy to check that the topological space axioms hold. □

This is an appropriate point at which to note that in general the same set can have different topologies on it. For example, if X contains at least two distinct points, then the discrete and the indiscrete topologies on X are different.

Definition 7.6 *Given two topologies $\mathcal{T}_1$ and $\mathcal{T}_2$ on the same set, we say that* $\mathcal{T}_1$ is coarser than $\mathcal{T}_2$ *if* $\mathcal{T}_1 \subseteq \mathcal{T}_2$.

It would be more accurate but also clumsier to say 'at least as coarse as' in place of 'coarser than'. The opposite of 'coarser' is 'finer': we say $\mathcal{T}_2$ *is finer than* $\mathcal{T}_1$ iff $\mathcal{T}_1$ is coarser than $\mathcal{T}_2$.

Remark For any topological space $(X, \mathcal{T})$, the indiscrete topology on X is coarser than $\mathcal{T}$ which in turn is coarser than the discrete topology on X.

Example 7.7 The *Sierpinski space* $\mathbb{S}$ consists of two points $\{0, 1\}$ with the topology $\{\emptyset, \{1\}, \{0, 1\}\}$. The topology of the Sierpinski space is finer than the indiscrete topology $\{\emptyset, \{0, 1\}\}$ on $\{0, 1\}$ but coarser than the discrete topology $\{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$ on $\{0, 1\}$.

Example 7.8 We can look at different ways of putting a topology on the set $S = \{0, 1, 2\}$. For example, let $\mathcal{T}_1 = \{\emptyset, S, \{0\}, \{0, 1\}\}$. It is easy to check 'by hand' (i.e. checking the axioms case-by-case) that this is a topology. Similarly, we may let $\mathcal{T}_2 = \{\emptyset, S, \{0\}, \{1, 2\}\}$, and again this is a topology. On the other hand, the collection $\{\emptyset, S, \{0, 1\}, \{1, 2\}\}$ is not a topology for S—we need to add in the intersection $\{1\}$ of $\{0, 1\}$ and $\{1, 2\}$ to make it up to a topology. In Exercise 7.1(b), you are asked to explore further the topologies that can be put on a set of three points.

We are now going to look at one more example of a topology.

Example 7.9 Let X be any non-empty set. The *co-finite* topology on X consists of the empty set together with every subset U of X such that $X \setminus U$ is finite.

Exercise 7.5 asks you to prove that this is indeed a topology. Note that when X is finite the co-finite topology is just the discrete topology; the interest in Example 7.9 is when X is infinite. In that case we shall see later that the resulting topological space is not metrizable.

In addition to these examples, there are two general ways of getting new topological spaces from old, which are explored in Chapter 10.

Exercise 7.1 List all possible topologies on (a) the set $\{0, 1\}$ and (b) the set $\{0, 1, 2\}$.

Exercise 7.2 Give an example of two topologies $\mathcal{T}_1$, $\mathcal{T}_2$ on the same set such that neither contains the other.

Exercise 7.3 Show that the intersection of two topologies on the same set X is also a topology on X, but that their union may or may not be a topology. Does the first result extend to the intersection of an arbitrary family of topologies on X?

Exercise 7.4 Prove that we get a topology for $\mathbb{N} = \{1, 2, 3, \ldots\}$ by taking the open sets to be $\emptyset$, $\mathbb{N}$ and $\{1, 2, \ldots, n\}$ for each $n \in \mathbb{N}$.

Exercise 7.5 Prove that for any set X the co-finite topology defined in Example 7.9 does give a topology for X. Show also that if X is finite then the co-finite topology is the discrete topology.

Exercise 7.6 Let $\mathcal{T}$ be the collection of all subsets of $\mathbb{R}$ consisting of $\emptyset$, $\mathbb{R}$ together with all intervals of the form $(-\infty, b)$. Show that $\mathcal{T}$ is a topology for $\mathbb{R}$.

8 Continuity in topological spaces; bases

Now we come to the main point of topological spaces as far as analysis is concerned: the definition of continuity. The second part of this chapter concerns bases, which are useful in considering continuity.

Definition

Bearing in mind how continuity in metric spaces can be characterized in terms of open sets, we define continuity as follows.

Definition 8.1 *We say that a map $f : X \to Y$ of topological spaces $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$ is* continuous *if $U \in \mathcal{T}_Y \Rightarrow f^{-1}(U) \in \mathcal{T}_X$. If necessary for clarity we say that f is $(\mathcal{T}_X, \mathcal{T}_Y)$-continuous.*

It is important to note that this definition concerns *inverse* images of open sets in Y and makes no reference to direct images of open sets in X.

For metric spaces we first defined continuity *at a point*. We can still do that, although at the present level of generality Definition 8.1 is the one most commonly used.

Definition 8.2 *With the notation of Definition 8.1, we say that f* is continuous at a point *$x \in X$ if, given any $U' \in \mathcal{T}_Y$ such that $f(x) \in U'$, there is some $U \in \mathcal{T}_X$ such that $x \in U$ and $f(U) \subseteq U'$.*

Then one can prove (see Exercise 8.2) that f is continuous iff it is continuous at every point of X.

Needless to say, Definition 8.1 does not conflict with our previous Definition 5.3(b) when X and Y are metric spaces; the next result states this.

Proposition 8.3 *If (X, d_X), (Y, d_Y) are metric spaces whose underlying topological spaces are $(X, \mathcal{T}_X)$, $(Y, \mathcal{T}_Y)$, then a map $f : X \to Y$ is (d_X, d_Y)-continuous iff it is $(\mathcal{T}_X, \mathcal{T}_Y)$-continuous.*

Proof This is just a translation of Proposition 5.37. □

We shall not prove much more about continuity before introducing further conditions on topological spaces, but there are a few results that can and should be proved at the present level of generality.

Proposition 8.4 *Given spaces $(X, \mathcal{T}_X)$, $(Y, \mathcal{T}_Y)$, $(Z, \mathcal{T}_Z)$ and continuous maps $f : X \to Y$, $g : Y \to Z$, the composition $g \circ f : X \to Z$ is continuous (more precisely, if f is $(\mathcal{T}_X, \mathcal{T}_Y)$-continuous and g is $(\mathcal{T}_Y, \mathcal{T}_Z)$-continuous then $g \circ f$ is $(\mathcal{T}_X, \mathcal{T}_Z)$-continuous).*

Proof Suppose that $U \subseteq Z$ is open in Z. Then since g is continuous, $g^{-1}(U)$ is open in Y, and hence since f is continuous, $f^{-1}(g^{-1}(U))$ is open in X. This says $(g \circ f)^{-1}(U)$ is open in X. Hence $g \circ f$ is continuous. □

Example 8.5 If f, g are real-valued functions of a real variable such that f is continuous on $[a, b]$ and g is continuous on some subset of $\mathbb{R}$ containing $f([a, b])$ then $x \mapsto g(f(x))$ is continuous on $[a, b]$.

Proposition 8.6 (*a*) *the identity map of any topological space is continuous*;

(*b*) *any constant map is continuous*;

(*c*) *if $\mathcal{T}_X$ is the discrete topology then any map $f : X \to Y$ to another topological space $(Y, \mathcal{T}_Y)$ is continuous*;

(*d*) *if $\mathcal{T}_Y$ is the indiscrete topology then any map $f : X \to Y$ from another topological space $(X, \mathcal{T}_X)$ is continuous.*

Proof This is Exercise 8.1. □

Homeomorphisms

The next definition is fundamental for topology: it tells you when two spaces are equivalent in the topological sense.

Definition 8.7 *A* homeomorphism *between topological spaces X and Y is a bijective map $f : X \to Y$ such that f and its inverse function f^{-1} are both continuous.*

Notice then that $U \subseteq X$ is open in X iff $f(U)$ is open in Y. So a homeomorphism is a one one correspondence that preserves all the structure that there is in a topological space, namely the open sets. (►The analogous notion for algebraic structures such as groups or vector spaces is an isomorphism.◄) If a homeomorphism exists between spaces we say that they are *homeomorphic* or just *equivalent*. One can check that this gives an equivalence relation on topological spaces (see Exercise 8.4). In topology,

we are interested in those properties of spaces, or quantities associated with spaces, which are preserved under all homeomorphisms; such properties or quantites are called *topological invariants*. We shall draw attention to some of these as we come to them. The study of homeomorphisms touches on the more geometric aspects of topology that appear in popular presentations of the subject. (For example, an overcoat is homeomorphic to a disc with two holes in it.)

Example 8.8 (a) Any two open intervals (a, b) and (c, d) in $\mathbb{R}$ with the topologies arising from the Euclidean metric are homeomorphic. As usual we assume here that $b > a$, $d > c$. A suitable homeomorphism $f : (a, b) \to (c, d)$ is given by

$$f(x) = c + \frac{(d-c)(x-a)}{b-a}.$$

(b) Any open interval (a, b) is homeomorphic to $\mathbb{R}$. By (a) it is enough to show that $(-1, 1)$ and $\mathbb{R}$ are homeomorphic. We may define a suitable homeomorphism $f : (-1, 1) \to \mathbb{R}$ by

$$f(x) = \frac{x}{1-|x|}.$$

Remark It is often easier to 'see' that two spaces are homeomorphic than to prove this explicitly; for example, think about a doughnut and a coffee cup.

Bases

In a metric space, any open set is a union of open balls, where in general infinitely many balls are involved in the union. (This was Exercise 5.13.) In a topological space it is often convenient to have some subfamily $\mathcal{B}$ of the open sets such that any open set is a union of sets from $\mathcal{B}$.

Definition 8.9 *Given a topological space $(X, \mathcal{T})$, a* basis *for $\mathcal{T}$ is a subfamily $\mathcal{B} \subseteq \mathcal{T}$ such that every set in $\mathcal{T}$ is a union of sets from $\mathcal{B}$.*

Example 8.10 In $\mathbb{R}^2$ with the Euclidean topology, the family

$$\{B_\varepsilon(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^2, \varepsilon > 0\} \quad \text{is a basis.}$$

Proof This is by *definition* of 'open set' in the Euclidean metric, together with Example 7.3 explaining how a topology is induced by a metric. □

Example 8.11 Let $S = \{0, 1, 2\}$. Then $\{\emptyset, S, \{0\}, \{1\}, \{0, 1\}\}$ is a topology for S. A basis for this topology is $\mathcal{B} = \{S, \{0\}, \{1\}\}$. (Note that we allow ourselves to take the union of *no* sets from $\mathcal{B}$, to get the empty set.)

Here is an application of basis.

Proposition 8.12 *To check that a map $f : X \to Y$ of topological spaces $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$ is continuous, it is enough to check that for each open set B in some basis for $\mathcal{T}_Y$, the inverse image $f^{-1}(B)$ is open in X.*

Proof For suppose it has been proved that $f^{-1}(B)$ is open in X for every B in some basis $\mathcal{B}$ for $\mathcal{T}_Y$. Any open set U in Y is a union $\bigcup_{i \in I} B_i$ for some indexing set I, where each B_i is in $\mathcal{B}$. So

$$f^{-1}(U) = f^{-1}(\bigcup_{i \in I} B_i) = \bigcup_{i \in I} f^{-1}(B_i).$$

This last set is a union of sets known to be open in X, hence it is open in X. Hence f is continuous. □

This application indicates that it may be worthwhile looking for 'economical' bases, that is to say bases with relatively few sets in them.

Example 8.13 Let $S = \{(x, y) \in \mathbb{R}^2 : x, y \in \mathbb{Q}\}$. Let $\mathcal{B}$ be the family of all open balls $B_q(x, y)$ as (x, y) ranges over S and q ranges over all positive rational numbers.

This gives a countable basis for the Euclidean topology on $\mathbb{R}^2$ (see Exercise 8.7). A topological space which admits a countable basis for open sets is called *second countable*. We shall not study second countable spaces in this book, but just note that they have potential for allowing inductive arguments.

There is another concept similar to but more general than bases, namely sub-bases. There is also another aspect of bases suppressed here since it can be confusing. Both of these are treated briefly on the web site.

► If you have met the idea of a basis in a vector space, notice that it is rather similar to a basis in a topological space: a basis for a vector space is a subset $\mathcal{B}$ of the vectors such that any vector can be expressed as a (finite) linear combination of vectors in $\mathcal{B}$. ◄

Exercise 8.1 Prove Proposition 8.6, that a map $f : X \to Y$ of topological spaces is continuous in each of the following cases:

(a) X and Y are the same space and f is the identity map;

(b) f is a constant map;

(c) the topology on X is discrete;

(d) the topology on Y is indiscrete.

Exercise 8.2 Prove that a map $f : X \to Y$ of topological spaces is continuous iff it is continuous at every point of X.

Exercise 8.3 Let $\mathbb{S}$ be the Sierpinski space of Example 7.7. For any subset A of a topological space X let $\chi_A : X \to \{0, 1\}$ be its *characteristic* or *indicator* function defined by

$$\chi_A = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

Prove that A is open in X iff $\chi_A : X \to \mathbb{S}$ is continuous.

Exercise 8.4 Prove that equivalence of topological spaces is an equivalence relation.

Exercise 8.5 Prove that the set of all open intervals $\{(a, b) : a, b \in \mathbb{R}, a < b\}$ is a basis for the usual topology on $\mathbb{R}$.

Exercise 8.6 Prove that a real-valued function $f : X \to \mathbb{R}$ on a space X is continuous if for any $x \in \mathbb{R}$ the sets $f^{-1}(x, \infty)$ and $f^{-1}(-\infty, x)$ are both open in X.
[Hint: note that $(a, b) = (-\infty, b) \cap (a, \infty)$.]

Exercise 8.7* Let $S = \{(x, y) \in \mathbb{R}^2 : x, y \in \mathbb{Q}\}$. Let $\mathcal{B}$ be the family of all open balls $B_q(x, y)$ as (x, y) ranges over S and q ranges over all strictly positive rational numbers. Prove that $\mathcal{B}$ is a countable basis for the Euclidean topology of $\mathbb{R}^2$.

Can you generalize this to $\mathbb{R}^n$?

9 Some concepts in topological spaces

Just as in metric spaces, in topological spaces there are notions of closed sets, closure, interior, and boundary. In fact, this chapter supersedes the discussion of these concepts in metric spaces. Alternatively, one may view the latter as an introduction to this chapter. Certainly, many of the results about closed subsets in metric spaces generalize to similar results in topological spaces. It is a good idea to compare proofs; sometimes the proof in this more general context is actually simpler.

Definition 9.1 *Let $(X, \mathcal{T})$ be a topological space. A subset V of X is said to be* closed in X *if $X \setminus V$ is open in X.*

The first example is immediate from the similarity between Definitions 6.1 and 9.1.

Example 9.2 If V is a closed subset of a metric space (X, d) in the sense of Definition 6.1 then it is closed in the underlying topological space $(X, \mathcal{T}_d)$.

This immediately gives specific examples coming from Example 6.2.

Example 9.3 In a space X with the co-finite topology (see Example 7.9) any finite subset is closed, since its complement is open.

The next proposition holds since a closed set is one whose complement is open.

Proposition 9.4 *Let X be a topological space. Then*

(C1) *$X, \emptyset$ are closed in X;*

(C2) *if V_1, V_2 are closed in X then $V_1 \cup V_2$ is closed in X;*

(C3) *if V_i is closed in X for all $i \in I$ then $\bigcap_{i \in I} V_i$ is closed in X.*

Proof To prove these we just take complements and apply (T1), (T2), and (T3). (Remember that taking complements interchanges union and intersection.) □

Similarly, we can express continuity in terms of closed sets.

Proposition 9.5 *A map $f : X \to Y$ of topological spaces is continuous iff $f^{-1}(V)$ is closed in X whenever V is closed in Y.*

Proof Again we just take complements and apply Definition 8.1. □

Just as in metric spaces, subsets of a topological space X are 'nothing like doors': most sets are neither open nor closed, and some sets are both open and closed (e.g. X and $\emptyset$). In particular, we need to beware that in order to show that a set is closed it is not enough to show that it is not open.

However, although not all sets in a space are closed, we can make up any set to a closed set called its *closure* by adding on any points which intuitively are very close to, though not actually in, the set. Similarly, we can get from any set to a related open set, called its *interior*. We shall explore closure in detail and just outline interior (and boundary), leaving some of their treatment to the exercises.

Definition 9.6 *A point a is a* point of closure of a subset A in a topological space X *if $U \cap A \neq \emptyset$ for any open subset U of X with $a \in U$. The* closure $\overline{A}$ of A in X *is the set of points of closure of A in X.*

Example 9.7 Given a subset A of a metric space (X, d), the closure $\overline{A}$ of A in the underlying topological space $(X, \mathcal{T}_d)$ is precisely the same as the closure of A in the metric space (X, d) as in Definition 6.7. This follows from the similarity of Definitions 9.6 and 6.7.

Example 9.8 In an infinite space X with the co-finite topology, the closure of any finite subset is itself while the closure of any infinite subset is X.

This follows from Exercise 9.4.

Just as in metric spaces, closure gives rise to an important definition.

Definition 9.9 *A subset A of a space X is* dense in X *if $\overline{A} = X$.*

Also as in metric spaces we can prove the following properties of closure.

Proposition 9.10 *Let A, B be subsets of a topological space X. Then*

(a) $A \subseteq \overline{A}$;
(b) *$A \subseteq B$ implies that $\overline{A} \subseteq \overline{B}$;*
(c) *A is closed in X iff $\overline{A} = A$;*
(d) $\overline{\overline{A}} = \overline{A}$;
(e) *$\overline{A}$ is closed in X;*
(f) *$\overline{A}$ is the smallest closed subset of X containing A.*

Proof Properties (a) and (b) are clear after a little thought—they follow from Definition 9.6. To prove (c), suppose first that A is closed in X. We shall show that no point of its complement is in $\overline{A}$. For if $x \in X \setminus A$ then $X \setminus A$ is an open in X containing x and $(X \setminus A) \cap A = \emptyset$. This shows that $x \notin \overline{A}$. Thus $\overline{A} \subseteq A$, and since $A \subseteq \overline{A}$ by (a), we get $\overline{A} = A$.

Conversely, if $\overline{A} = A$ we shall show that $X \setminus A$ is open in X, hence A is closed in X. For if $x \in X \setminus A$ then $x \notin \overline{A}$, so there is some open set, say U_x with $x \in U_x$ and $U_x \cap A = \emptyset$, whence $U_x \subseteq X \setminus A$. So $X \setminus A$ is open by Proposition 7.2.

To prove (d), first note that $\overline{A} \subseteq \overline{\overline{A}}$ by (b). Now let $x \in \overline{\overline{A}}$ and let U be any open subset of X containing x. Then $U \cap \overline{A} \neq \emptyset$, say $y \in U \cap \overline{A}$. Then U is an open set containing the point $y \in \overline{A}$ so $U \cap A \neq \emptyset$. This proves that $x \in \overline{A}$, hence $\overline{\overline{A}} \subseteq \overline{A}$. So $\overline{\overline{A}} = \overline{A}$.

Now (c) follows from (c) and (d). For $\overline{\overline{A}} = \overline{A}$ by (d), so by (c) applied with A replaced by $\overline{A}$ we get that $\overline{A}$ is closed in X.

Finally, we expand on what (f) means and then prove it. It means that $A \subseteq \overline{A}$ (which we know already from (a)), and also that if B is any closed subset of X satisfying $A \subseteq B$ then $\overline{A} \subseteq B$. So suppose B is such a closed set. By (b) we have $\overline{A} \subseteq \overline{B}$. But by (c) we have $\overline{B} = B$. So $\overline{A} \subseteq B$ as required. □

Another way of expressing (f) is to say that $\overline{A}$ is the intersection, call it V, of all those sets which are closed in X and contain A. For by (C3), V is also closed in X, and V contains A, so by the argument in the previous paragraph we have $\overline{A} \subseteq V$. But $\overline{A}$ is itself closed in X by (e), and contains A by (a), so it is one of the sets we take the intersection of to form V, hence $V \subseteq \overline{A}$. This proves that $V = \overline{A}$.

Again as for metric spaces we can express continuity of maps of topological spaces in terms of closure.

Proposition 9.11 *A map $f : X \to Y$ of topological spaces is continuous iff $f(\overline{A}) \subseteq \overline{f(A)}$ for every $A \subseteq X$.*

Proof This is Exercise 9.7. □

There are analogues of Proposition 9.4 for closure.

Proposition 9.12 *Let $A_1, A_2, \ldots, A_m$ be subsets of a topological space X. Then*

$$\overline{\bigcup_{i=1}^{m} A_i} = \bigcup_{i=1}^{m} \overline{A_i}.$$

Proposition 9.13 *For each i in some indexing set I, let A_i be a subset of the topological space X. Then*

$$\overline{\bigcap_{i\in I} A_i} \subseteq \bigcap_{i\in I} \overline{A_i}.$$

Proof The proofs of Propositions 9.12 and 9.13 are Exercise 9.6. □

Equality does not necessarily hold in Proposition 9.13 even when the indexing set is finite—we have already seen examples of this in the case of metric spaces.

As in metric spaces there are two additional concepts that we mention more briefly: interior and boundary of a subset in a topological space. The interior is a 'dual' concept to closure.

Definition 9.14 *A point a is an interior point of a subset A in a topological space X if there exists some set U which is open in X and with $a \in U \subseteq A$. The set of all interior points of A is called* the interior of A, *written $\mathring{A}$.*

Example 9.15 The interiors of the intervals $[a, b]$, $[a, b)$, $(a, b]$, (a, b) in $\mathbb{R}$ are all the same, namely (a, b). The interior of $\mathbb{Q}$ in $\mathbb{R}$ is $\emptyset$.

We relate interior to closure in the following result.

Proposition 9.16 *We have $\overline{X \setminus A} = X \setminus \mathring{A}$ for any subset A of a space X.*

Proof Let $x \in X$. Then x is in $\overline{X \setminus A}$ iff any set U open in X and containing x has non-empty intersection with $X \setminus A$. But this is true iff no set U open in X and containing x is contained in A, which is true iff x is not in $\mathring{A}$. □

Armed with this result, we can deduce results about interior which are dual to those already proved about closure, as in the following analogue of Proposition 9.10.

Proposition 9.17 *Let A, B be subsets of a space X. Then*

(i) $\mathring{A} \subseteq A$;

(ii) *if $A \subseteq B$ then $\mathring{A} \subseteq \mathring{B}$;*

(iii) *A is open in X iff $\mathring{A} = A$;*

(iv) $\mathring{\mathring{A}} = \mathring{A}$;

(v) $\mathring{A}$ *is open in* X;

(vi) $\mathring{A}$ *is the largest open set of* X *contained in* A.

Proof These may be proved either from scratch or deduced from Proposition 9.10. As a sample we deduce (iii) from Proposition 9.10(iii); the other parts are Exercise 9.9. By definition A is open in X iff $X \setminus A$ is closed in X. By Proposition 9.10(iii), $X \setminus A$ is closed in X iff $\overline{X \setminus A} = X \setminus A$, and by Proposition 9.16, the latter holds iff $X \setminus \mathring{A} = X \setminus A$, which is true iff $\mathring{A} = A$. □

Results dual to Proposition 9.11, Proposition 9.12, and Proposition 9.13 are in the exercises at the end of this chapter. The first of these generalizes Exercise 6.20. The other two could have been stated for metric spaces in Chapter 6 but have been 'saved' until now.

The boundary (or *frontier*) of a set is defined as follows.

Definition 9.18 *The* boundary ∂A *of a subset* A *of a space* X *is the set* $\overline{A} \setminus \mathring{A}$.

Example 9.19 The boundary of each interval $[a, b]$, $[a, b)$, $(a, b]$, (a, b) in $\mathbb{R}$ is $\{a, b\}$. The boundary of $\mathbb{Q}$ in $\mathbb{R}$ is $\mathbb{R}$.

The definition of boundary can be given more symmetrically.

Proposition 9.20 *The boundary of a subset* A *in a space* X *is* $\overline{A} \cap \overline{X \setminus A}$.

Proof In general, if C, D are subsets of a set X then $D \setminus C = (X \setminus C) \cap D$ (see Exercise 2.1); we apply this with C, D taken to be $\mathring{A}$, $\overline{A}$, respectively. Using also Proposition 9.16 we have

$$\partial A = \overline{A} \setminus \mathring{A} = (X \setminus \mathring{A}) \cap \overline{A} = \overline{X \setminus A} \cap \overline{A}.$$

□

This more symmetric approach immediately gives the following corollary.

Corollary 9.21 *We have* $\partial A = \partial(X \setminus A)$ *for any subset* A *of a space* X.

Remark Given a subset A of a metric space (X, d), the interior $\mathring{A}$ of A in the underlying topological space $(X, \mathcal{T}_d)$ is precisely the same as the interior of A in the metric space (X, d); similarly, the boundary ∂A of A in the underlying topological space $(X, \mathcal{T}_d)$ is precisely the same as the boundary of A in the metric space (X, d).

We end this chapter with a brief mention of 'neighbourhoods'. These are much used in more advanced textbooks. There is no unanimous agreement about the definition; here is the more common one.

Definition 9.22 *A* neighbourhood *of a point x in a space X is a subset N of X which contains an open subset of X containing x.*

(Some writers insist that a neighbourhood itself should be open in X.) This concept is particularly useful in discussing the nature of the topology 'around a point' in a space.

Exercise 9.1 Prove that any subset of a discrete topological space X is closed in X.

Exercise 9.2 Let X be an infinite set with the co-finite topology. Determine which subsets of X are closed in X.

Exercise 9.3 Find open sets $U, V \subseteq \mathbb{R}$ such that $U \cap \overline{V}$, $\overline{U} \cap V$, $\overline{U} \cap \overline{V}$, $\overline{U \cap V}$ are all distinct.

Exercise 9.4 Suppose that X is an infinite set with the co-finite topology, and $A \subseteq X$. Find $\overline{A}$ in the cases (a) A finite and (b) A infinite.

Exercise 9.5 Give either a proof of, or a counterexample to, each of the following:

(a) If $f : X \to Y$ is a continuous map of topological spaces and A is a closed subset of X then $f(A)$ is a closed subset of Y

(b) If A is open in a topological space X and $B \subseteq X$ then $A \cap \overline{B} = \overline{A \cap B}$.

(c) If $f : X \to Y$ is a continuous map of topological spaces and $B \subseteq Y$ then $f^{-1}(\overline{B}) = \overline{f^{-1}(B)}$.

Exercise 9.6 (a) Let $A_1, A_2, \ldots, A_m$ be subsets of a topological space X. Prove that

$$\overline{\bigcup_{i=1}^{m} A_i} = \bigcup_{i=1}^{m} \overline{A_i}$$

(b) For each i in some indexing set I, let A_i be a subset of the topological space X Prove that

$$\overline{\bigcap_{i\in I} A_i} \subseteq \bigcap_{i\in I} \overline{A_i}.$$

Exercise 9.7 Prove that a map $f : X \to Y$ of topological spaces is continuous iff $f(\overline{A}) \subseteq \overline{f(A)}$ for every $A \subseteq X$

Exercise 9.8 Suppose that X is an infinite set with the co-finite topology and $A \subseteq X$
Find the interior $\overset{\circ}{A}$ in the cases (a) A is finite and (b) A is infinite. Find also the boundary ∂A in each case.

Exercise 9.9 Let A, B be subsets of a space X. Prove

(a) $\overset{\circ}{A} \subseteq A$;

(b) if $A \subseteq B$ then $\overset{\circ}{A} \subseteq \overset{\circ}{B}$;

(c) A is open in X iff $\overset{\circ}{A} = A$;

(d) $\overset{\circ}{\overset{\circ}{A}} = \overset{\circ}{A}$;

(e) $\overset{\circ}{A}$ is open in X;

(f) $\overset{\circ}{A}$ is the largest open set of X contained in A.

Exercise 9.10 Prove that a map $f : X \to Y$ of topological spaces is continuous iff for every subset $B \subseteq Y$ we have that $f^{-1}\left(\overset{\circ}{B}\right)$ is contained in the interior of $f^{-1}(B)$.

Exercise 9.11 Let $A_1, A_2, \ldots, A_m$ be subsets of a topological space X. Prove that the interior of $\bigcap_{i=1}^{m} A_i$ equals $\bigcap_{i=1}^{m} \overset{\circ}{A}_i$.

Exercise 9.12 Let X be a topological space and suppose that for each i in some indexing set I we are given a subset A_i of X Prove that $\bigcup_{i\in I} \overset{\circ}{A}_i$ is contained in the interior of $\bigcup_{i\in I} A_i$. Give an example to show that this containment may be strict even when there are only two sets involved in the union

Exercise 9.13 Suppose that A is a subset of a topological space X. Prove that the boundary ∂A is closed in X.

Exercise 9.14 Suppose that A is a subset of a topological space X. Prove that

(a) A is closed in X iff $\partial A \subseteq A$.

(b) $\partial A = \emptyset$ iff A is open and closed in X.

Exercise 9.15 Let A be a subspace of a topological space X. Show that $\overline{A}$ is the disjoint union of $\mathring{A}$ and ∂A. Deduce that if B is another subspace of X such that $B \cap A \neq \emptyset$ then either $B \cap \partial A \neq \emptyset$ or $B \cap \mathring{A} \neq \emptyset$.

Exercise 9.16 Suppose that A is a non-empty subset of a topological space X. Prove that X is the union of three mutually disjoint subsets $\mathring{A}$, ∂A, and the interior of $X \setminus A$.

10 Subspaces and product spaces

In this chapter we consider two ways of getting new spaces from old—subspaces and product spaces.

Example 10.1 The interval $[a, b]$ in the real line, with the topology arising from the usual metric, is a topological subspace of $\mathbb{R}$ with its usual topology.

Example 10.2 The Euclidean plane with its usual topology is the topological product of the Euclidean line with itself.

These concepts generalize the analogous concepts in metric spaces, as we shall check later, but initially we introduce them in the topological context.

Subspaces

Definition 10.3 *Let $(X, \mathcal{T})$ be a topological space and let A be a non-empty subset of X. The* subspace *topology on A is $\mathcal{T}_A = \{A \cap U : U \in \mathcal{T}\}$.*

Other names for this topology are the *induced* topology and the *relative* topology. It is straightforward to check that $\mathcal{T}_A$ is indeed a topology for A (Exercise 10.2). We call A with this topology a (topological) *subspace* of X. Since we loosely use the same name for a topological space and its set of points, A is referred to as either a subset or a subspace according to the context; in the latter case it is always assumed to have the subspace topology. When it is desirable for clarity we use the full notation $(A, \mathcal{T}_A)$ for this subspace.

The next few results are intended to help explain the above choice for a subspace topology.

Proposition 10.4 *Let $(X, \mathcal{T})$ be a topological space and let A be a non-empty subset of X with the subspace topology $\mathcal{T}_A$. Then the inclusion map $i : A \to X$ defined by $i(a) = a$ for all $a \in A$, is $(\mathcal{T}_A, \mathcal{T})$-continuous.*

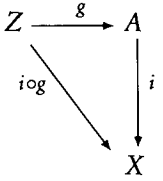

Figure 10.1. Subspace topologies

Proof For any subset $U \subseteq X$ we have $i^{-1}(U) = U \cap A$. In particular, if $U \in \mathcal{T}$ then $i^{-1}(U) = U \cap A$ is in $\mathcal{T}_A$, so i is $(\mathcal{T}_A, \mathcal{T})$-continuous. □

Corollary 10.5 *Let $f : X \to Y$ be a continuous map of topological spaces $(X, \mathcal{T})$, $(Y, \mathcal{T}')$ and let A be a non-empty subset of X with the subspace topology $\mathcal{T}_A$. Then the restriction $f|_A : A \to Y$ is $(\mathcal{T}_A, \mathcal{T}')$-continuous.*

Proof This follows from 10.4 since $f|_A = f \circ i$, and a composition of continuous maps is continuous (Proposition 8.4). □

The next result explores the reason for our choice of subspace topology further. In it, for simpicity, we drop names for the topologies.

Proposition 10.6 *Let X be a topological space, let A be a subspace of X and let $i : A \to X$ be the inclusion map. Suppose that Z is a topological space and that $g : Z \to A$ is a map. Then g is continuous iff $i \circ g : Z \to X$ is continuous.*

Proof Figure 10.1 may be helpful in following this proof. If g is continuous then so is the composition $i \circ g$ since i is continuous by 10.4.

Conversely suppose that $i \circ g$ is continuous. Then for any open subset V of A, we know $V = U \cap A$ for some U open in X. Hence

$$g^{-1}(V) = g^{-1}(U \cap A) = g^{-1}(i^{-1}(U)) = (i \circ g)^{-1}(U)$$

which is open in Z by continuity of $i \circ g$. So g is continuous. □

Example 10.7 Let (X, d) be a metric space and A a non-empty subset of X. Let $\mathcal{T} = \mathcal{T}_d$ be the topology on X arising from the metric d, and let d_A be the subspace metric on A (see Definition 5.8). Then the subspace topology $\mathcal{T}_A$ on A coincides with the topology arising from the metric d_A.

Proof The proof is Exercise 10.4. □

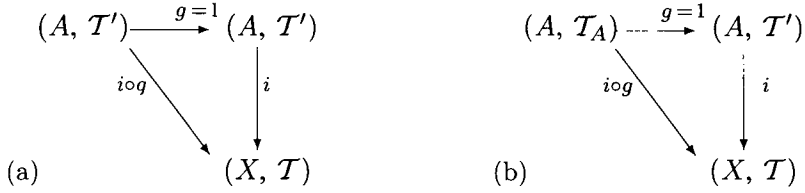

Figure 10.2. Proof of Proposition 10.8

The exercises at the end of this chapter include ways in which the idea of a subspace interacts with other concepts we have explored such as closed sets and closure. Our final result about subspaces is intended to show that the choice of subspace topology is 'inevitable'. It is a little more sophisticated than the results above.

Proposition 10.8 *With notation as in Proposition 10.6, the subspace topology $\mathcal{T}_A$ on A is the only topology satisfying Proposition 10.6 for all possible maps g.*

Proof Suppose that $\mathcal{T}'$ is a topology on A such that for any topological space $(Z, \mathcal{T}_Z)$ and any map $g : Z \to A$ we have g is $(\mathcal{T}_Z, \mathcal{T}')$-continuous iff $i \circ g$ is $(\mathcal{T}_Z, \mathcal{T})$-continuous, where as usual $i : A \to X$ is the inclusion. We wish to show that $\mathcal{T}' = \mathcal{T}_A$. Figure 10.2 may help in following the proof.

First take $Z = A$, $\mathcal{T}_Z = \mathcal{T}'$, and g the identity map of A, as in Figure 10.2(a). Then g is certainly $(\mathcal{T}', \mathcal{T}')$-continuous, so i, which is the same as $i \circ g$, is $(\mathcal{T}', \mathcal{T})$-continuous. Hence for any $U \in \mathcal{T}$ we have $i^{-1}(U) \in \mathcal{T}'$. But any $V \in \mathcal{T}_A$ is of the form $U \cap A$ or equivalently $i^{-1}(U)$ for some $U \in \mathcal{T}$, so we get $\mathcal{T}_A \subseteq \mathcal{T}'$.

Secondly take $Z = A$, $\mathcal{T}_Z = \mathcal{T}_A$, g the identity map of A, as in Figure 10.2(b). Since $i \circ g$ is the same as i, which is is $(\mathcal{T}_A, \mathcal{T})$-continuous, g is $(\mathcal{T}_A, \mathcal{T}')$-continuous, which tells us that any $V \in \mathcal{T}'$ is also in $\mathcal{T}_A$, in other words $\mathcal{T}' \subseteq \mathcal{T}_A$.

From these two inclusions we get $\mathcal{T}' = \mathcal{T}_A$ as required. □

Products

Given topological spaces $(X, \mathcal{T}_X)$, $(Y, \mathcal{T}_Y)$, we want to get a sensible topology for the set $X \times Y$. Here are two ways of motivating the choice.

First, when we write parametric equations for a curve in the plane, say

$$\left.\begin{aligned} x &= x(t) \\ y &= y(t) \end{aligned}\right\} \qquad a \leqslant t \leqslant b,$$

we expect the curve to be continuous as a function from $[a, b]$ to $\mathbb{R}^2$ provided the 'coordinate functions' $x(t)$ and $y(t)$ are continuous functions of t (from $[a, b]$ to $\mathbb{R}$). In elementary mathematics this is taken for granted, or else the curve is defined to be continuous when x and y are continuous functions of t. Later, the reader may have seen continuity of a function $f : [a, b] \to \mathbb{R}^2$ defined in terms of the Euclidean metrics on $[a, b]$ and $\mathbb{R}^2$, and it is then easy to show that f is continuous iff the first and second coordinate functions are continuous. These coordinate functions are the compositions $p_X \circ f$ and $p_Y \circ f$ where p_X, p_Y are the projections of the plane onto the x-axis and the y-axis, respectively.

More generally for any sets X, Y we can define projections 'on the axes' $p_X : X \times Y \to X$, $p_Y : X \times Y \to Y$ by $p_X(x, y) = x$, $p_Y(x, y) = y$ for any $(x, y) \in X \times Y$. Then a map $f : Z \to X \times Y$ from any set Z defines two 'coordinate maps' $p_X \circ f : Z \to X$, $p_Y \circ f : Z \to Y$. Conversely any two maps $g : Z \to X$, $h : Z \to Y$ give a map $f : Z \to X \times Y$ by the formula $f(z) = (g(z), h(z))$, and f has the property that its two coordinate functions are g, h. If we want to follow the analogy with the case of a parametrized curve in the plane, then when X and Y have topologies, it would be good to have a topology on $X \times Y$ such that f is continuous iff both g and h are continuous. It turns out that the topology in Proposition 10.9 below fits the bill.

Another way of motivating the choice of product topology is to consider the special case of $X = Y = \mathbb{R}$. How do we characterize open sets of $\mathbb{R}^2$ in terms of open sets in $X = \mathbb{R}$ and $Y = \mathbb{R}$? If U is an open subset of $\mathbb{R}^2$ and $x = (x_1, x_2) \in U$, then there exists $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq U$. So for any ε_1, ε_2 such that $0 < \varepsilon_1 \leqslant \varepsilon$ and $0 < \varepsilon_2 \leqslant \varepsilon$ the rectangle centred on x with sides of lengths $2\varepsilon_1$, $2\varepsilon_2$ parallel to the axes is contained in $B_\varepsilon(x)$ and hence in U, i.e. $(x_1 - \varepsilon_1, x_1 + \varepsilon_1) \times (x_2 - \varepsilon_2, x_2 + \varepsilon_2) \subseteq U$. This is true for every point x in U, so U is a union of such rectangular open sets. (In general the number of rectangles in the union will be vast - think of the case when U is an open disc in the plane, for example.) This suggests a description of the product topology in general.

Proposition 10.9 *Suppose that $(X, \mathcal{T}_X)$, $(Y, \mathcal{T}_Y)$ are topological spaces, and let $\mathcal{T}_{X\times Y}$ be the family of all unions of sets of the form $U \times V$ where*

$U \in \mathcal{T}_X$ *and* $V \in \mathcal{T}_Y$. *Then* $\mathcal{T}_{X\times Y}$ *is a topology for* $X \times Y$, *called* the product topology.

The space $(X \times Y, \mathcal{T}_{X\times Y})$ is called the *topological product* of $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$.

Proof of 10.9. We check that $\mathcal{T}_{X\times Y}$ is a topology.

(T1) You can get the empty set either by taking the union of *no* sets of the form $U \times V$ with $U \in \mathcal{T}_X$ and $V \in \mathcal{T}_Y$, or by taking $\emptyset \times \emptyset$ since $\emptyset \in \mathcal{T}_X$ and $\emptyset \in \mathcal{T}_X$. So $\emptyset \in \mathcal{T}_{X\times Y}$. Also, since $X \in \mathcal{T}_X$ and $Y \in \mathcal{T}_Y$, we have $X \times Y \in \mathcal{T}_{X\times Y}$.

(T2) Suppose that W_1 and W_2 are in $\mathcal{T}_{X\times Y}$, say

$$W_1 = \bigcup_{i\in I} U_{1i} \times V_{1i}, \quad W_2 = \bigcup_{j\in J} U_{2j} \times V_{2j} \text{ for some indexing sets } I, J,$$

where each $U_{1i}, U_{2j} \in \mathcal{T}_X$ and each $V_{1i}, V_{2j} \in \mathcal{T}_Y$. Then $W_1 \cap W_2$ is the union of all the possible intersections $(U_{1i} \times V_{1i}) \cap (U_{j2} \times V_{2j})$ (see Exercise 2.6). But (see Exercise 2.5)

$$(U_1 \times V_1) \cap (U_2 \times V_2) = (U_1 \cap U_2) \times (V_1 \cap V_2).$$

But $U_1, U_2 \in \mathcal{T}_X$ and $V_1, V_2 \in \mathcal{T}_Y$, so $U_1 \cap U_2 \in \mathcal{T}_X$ and $V_1 \cap V_2 \in \mathcal{T}_Y$. Thus $W_1 \cap W_2$ is in $\mathcal{T}_{X\times Y}$.

(T3) A union of unions of sets of the form $U \times V$ where $U \in \mathcal{T}_X$ and $V \in \mathcal{T}_Y$ is again a union of sets of this form, hence is in $\mathcal{T}_{X\times Y}$. □

From its definition, the product topology has basis

$$\mathcal{B} = \{U \times V : U \in \mathcal{T}_X, V \in \mathcal{T}_Y\}.$$

NB Although this basis is convenient, unfortunately its use encourages a common error. One should not assume that any open set in the product is a single 'rectangular open set' such as $U \times V$. In general an open set in $X \times Y$ will be the union of (possibly a very large number of) sets of the form $U \times V$ as above.

The next two results show that our choice of product topology does have the first property we were hoping for above; they are crucial to dealing with products. For a slightly sophisticated analysis (as in Proposition 10.8) showing that the product topology is the *unique* topology that satisfies these results, we refer to the web site.

Proposition 10.10 *With the notation of Proposition 10.9, the two projection maps $p_X : X \times Y \to X$ and $p_Y : X \times Y \to Y$ are continuous, where $p_X(x, y) = x$ and $p_Y(x, y) = y$ for all $(x, y) \in X \times Y$.*

Proof For any U open in X we have $p_X^{-1}(U) = U \times Y$, which is open in $X \times Y$, so p_X is continuous. Similarly p_Y is continuous. □

Proposition 10.11 *With the notation of Propositions 10.9 and 10.10, any map $f : Z \to X \times Y$ from a topological space Z into the topological product $X \times Y$ is continuous iff both $p_X \circ f : Z \to X$ and $p_Y \circ f : Z \to Y$ are continuous.*

Proof In one direction this is easy: if f is continuous then so are the compositions $p_X \circ f$ and $p_Y \circ f$, using Proposition 10.10.

Suppose now that $p_X \circ f$ and $p_Y \circ f$ are continuous. To show that f is continuous, we note that by Proposition 8.12 it is enough to show that $f^{-1}(B)$ is open in Z for any B in the basis $\mathcal{B}$ described above. So let U, V be open in X, Y, respectively. Then

$$U \times V = (U \times Y) \cap (X \times V) = p_X^{-1}(U) \cap p_Y^{-1}(V).$$

Since inverse image preserves intersections, it follows that

$$f^{-1}(U \times V) = (f^{-1}p_X^{-1}(U)) \cap (f^{-1}p_Y^{-1}(V)) = (p_X \circ f)^{-1}(U) \cap (p_Y \circ f)^{-1}(V)$$

which is open in Z as an intersection of two sets which are open by continuity of $p_X \circ f$ and $p_Y \circ f$. Hence f is continuous. □

Proposition 10.11 has many applications. Here are three examples.

Proposition 10.12 *If $f : X \to X'$ and $g : Y \to Y'$ are continuous, then so is $f \times g : X \times Y \to X' \times Y'$ defined by $(f \times g)(x, y) = (f(x), g(y))$.*

Proof We know that $p_{X'} \circ (f \times g) = f \circ p_X$, since for any $(x, y) \in X \times Y$,

$$p_{X'} \circ (f \times g)(x, y) = f(x) = f \circ p_X(x, y).$$

This kind of fact is readily seen in what is called a 'commutative diagram': in Figure 10.3 the compositions the two ways round the square are equal. Since both f and p_X are continuous, so is $f \circ p_X$. Hence $p_{X'} \circ (f \times g)$ is continuous. Similarly $p_{Y'} \circ (f \times g)$ is continuous, so by Proposition 10.11 $f \times g$ is continuous. □

This is less intricate than the proof for metric spaces (Proposition 5.19).

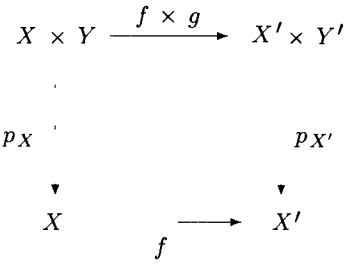

Figure 10.3. A commutative diagram

Proposition 10.13 *For any topological space X let $\Delta : X \to X \times X$ be the 'diagonal' map defined by $\Delta(x) = (x, x)$. Then Δ is continuous.*

Proof In this case, if p_1, p_2 are the projections of $X \times X$ on the first and second factors then both $p_1 \circ \Delta$ and $p_2 \circ \Delta$ are the identity map of X, hence continuous. The result follows from Proposition 10.11. □

Proposition 10.14 *Let X and Y be topological spaces, and let $y_0 \in Y$. Define $i_{y_0} : X \to X \times Y$ by $i_{y_0}(x) = (x, y_0)$. Then i_{y_0} is continuous.*

Proof The compositions $p_X \circ i_{y_0} : X \to X$ and $p_Y \circ i_{y_0} : X \to Y$ are respectively the identity function and the constant function with value y_0, so both are continuous and the result follows from Proposition 10.11. □

Just as in the case of metric spaces, we may form algebraic combinations of real-valued functions on a topological space.

Proposition 10.15 *If $f, g : X \to \mathbb{R}$ are continuous real-valued functions on a topological space X, then so are (a) $|f|$, (b) $f + g$, (c) fg. (d) If g is never zero on X then $1/g$ is also continuous.*

Proof This is exactly the same as the second proof of Proposition 5.17. For example, fg is the composition

$$X \xrightarrow{\Delta} X \times X \xrightarrow{f \times g} \mathbb{R} \times \mathbb{R} \xrightarrow{m} \mathbb{R},$$

where m is multiplication of real numbers. Since each of Δ, $f \times g$ and m is continuous so is fg. □

Example 10.16 Define $f : \mathbb{R}^2 \to \mathbb{R}^2$ by $f(x, y) = (x^2 + y^2, xy)$. Then f is continuous.

Proof Define $g, h : \mathbb{R}^2 \to \mathbb{R}$ by $g(x, y) = x^2 + y^2$ and $h(x, y) = xy$. By Proposition 10.11, it is enough to show that g, h are continuous. Since h is the product of the projections p_1, p_2 on the coordinates axes, it is continuous by Proposition 10.15 and Proposition 10.10. Similarly $(x, y) \mapsto x^2$ and $(x, y) \mapsto y^2$ are both continuous, so g is continuous by Proposition 10.15. □

In appropriate circumstances, the product topology is compatible with product metrics.

Proposition 10.17 *Let (X, d_X), (Y, d_Y) be metric spaces. Let $\mathcal{T}_X$ be the topology arising from d_X and $\mathcal{T}_Y$ the topology arising from d_Y. Let d denote any one of the product metrics on $X \times Y$ defined in Example 5.10, and let $\mathcal{T}_d$ be the topology on $X \times Y$ arising from d. Then $\mathcal{T}_d$ coincides with the product topology of the spaces $(X, \mathcal{T}_X)$, $(Y, \mathcal{T}_Y)$.*

Proof The proof is Exercise 10.14. □

Graphs

The product topology enables us to study graphs of maps.

Proposition 10.18 *Suppose that $f : X \to Y$ is a continuous map of topological spaces and that G_f is the graph of f, that is the subset of $X \times Y$ defined by $G_f = \{(x, y) \in X \times Y : f(x) = y\}$, with the topology induced by the product topology on $X \times Y$. Then the map $x \to (x, f(x))$ defines a homeomorphism θ from X to G_f.*

Proof The map $\phi : G_f \to X$ defined by $\phi(x, f(x)) = x$ is the set-theoretic inverse of θ, since both the composition $x \mapsto (x, f(x)) \mapsto x$ and the composition $(x, f(x)) \mapsto x \mapsto (x, f(x))$ are identity maps. Since ϕ is the restriction to the subspace G_f of the projection $p_X : X \times Y \to X$, it is continuous by Propositions 10.10 and 10.5. To see that θ is continuous we invoke Proposition 10.11: the compositions $p_X \circ \theta$ and $p_Y \circ \theta$ are respectively $x \mapsto x$ (the identity map of X) and $x \mapsto f(x)$ (the map f), and both of these are continuous . Thus ϕ and θ are mutually inverse homeomorphisms. □

Example 10.19 The real line and the parabolic subspace P of $\mathbb{R}^2$ given by $P = \{(x, x^2) : x \in \mathbb{R}\}$ are homeomorphic

This follows from Proposition 10.18 applied to the map $x \mapsto x^2$.

Postscript on products

The definition of product topology extends by induction to any finite number of factors. In particular, it follows from Proposition 10.17 and induction that the product topology on $\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R}$ (n factors) coincides with the topology arising from the Euclidean metric.

In fact one can define topological products of infinitely many topological spaces; that is beyond the scope of this book (but see for example Section 8 of Willard (2004)).

We end this chapter with a straightforward but useful criterion for a subset of a product to be open in the topological product.

Proposition 10.20 *With notation as in 10.9, $W \subseteq X \times Y$ is open in $X \times Y$ iff for any $(x, y) \in W$ there exist subsets U, V of X, Y respectively which are open in X, Y and with $(x, y) \in U \times V \subseteq W$.*

Proof Suppose that $W \in \mathcal{T}_{X \times Y}$. By definition, then W is a union of sets of the form $U \times V$ with $U \in \mathcal{T}_X$, $V \in \mathcal{T}_Y$. If $(x, y) \in W$ then (x, y) is in at least one such set $U \times V$, and then $(x, y) \in U \times V \subseteq W$.

Conversely if the criterion is satisfied, then for each $(x, y) \in W$ there exist $U_{(x,y)} \in \mathcal{T}_X$ and $V_{(x,y)} \in \mathcal{T}_Y$ such that $(x, y) \in U_{(x,y)} \times V_{(x,y)} \subseteq W$. It is easy to check that

$$W = \bigcup_{(x,y) \in W} U_{(x,y)} \times V_{(x,y)},$$

so W is the kind of union that qualifies to be in the product topology. □

Exercise 10.1 Let $X = \{a, b, c\}$ $\mathcal{T} = \{\emptyset, X, \{a, b\}, \{a, c\}, \{a\}\}$ $A = \{a, b\}$. Find the subspace topology $\mathcal{T}_A$ for A.

Exercise 10.2 Let $(X, \mathcal{T})$ be a topological space and let A be a non-empty subset of X. Prove that the family $\mathcal{T}_A = \{A \cap U : U \in \mathcal{T}\}$ is a topology for A.

Exercise 10.3 Suppose that A is a non-empty subset of a set X, and let $\mathcal{T}$ be the co-finite topology on X (see Example 7.9). Prove that $\mathcal{T}$ induces the co-finite topology on A.

Exercise 10.4 Given a metric subspace (A, d_A) of a metric space (X, d), let $\mathcal{T} = \mathcal{T}_d$ be the topology on X induced by d. Prove that the relative topology $\mathcal{T}_A$ induced on A by $\mathcal{T}$ coincides with the topology on A induced by the metric d_A.

Exercise 10.5 Suppose that $(A, \mathcal{T}_A)$ is a subspace of a space $(X, \mathcal{T})$ and that $V \subseteq X$ is closed in X. Prove that $V \cap A$ is closed in $(A, \mathcal{T}_A)$.
[Hint: use Exercise 2.2: $A \setminus (V \cap A) = A \cap (X \setminus V)$.]

Exercise 10.6 Suppose that $(A, \mathcal{T}_A)$ is a subspace of a topological space $(X, \mathcal{T})$ and let $W \subseteq A$.

(a) If W is open in A (that is, if $W \in \mathcal{T}_A$) and A is open in X (that is, if $A \in \mathcal{T}$) then W is open in X (that is, $W \in \mathcal{T}$).

(b) If W is closed in A and A is closed in X then W is closed in X.

Exercise 10.7 Suppose that $f : X \to Y$ is a map of topological spaces. Prove that f is continuous in each of the following cases.

(a) $X = \bigcup_{i \in I} U_i$ where $\{U_i \,.\, i \in I\}$ is a family of open subsets of X and $f|U_i$ is continuous for each $i \in I$.

(b) $X = \bigcup_{i=1}^{n} V_i$ where each V_i is a closed subset of X and $f|V_i$ is continuous for each i.

[Hint: Recall Exercise 3.13.]

Exercise 10.8 Given subsets $A \subseteq B$ of a topological space $(X, \mathcal{T})$, with $A \neq \emptyset$, let $\mathcal{T}_A$, $\mathcal{T}_B$ be the subspace topologies on A, B induced by $\mathcal{T}$. Prove that $\mathcal{T}_A$ coincides with the topology on A induced by $\mathcal{T}_B$.

Exercise 10.9 Let $A \subseteq X_1 \subset X_2$ where X_2 is a topological space, X_1 is a subspace of X_2 and A is a subset of X_1, and let B_i denote the closure of A in X_i for $i = 1, 2$. Prove that

(a) $B_1 = B_2 \cap X_1$,

(b) if X_1 is closed in X_2 then $B_1 = B_2$.

Exercise 10.10 Suppose that X, Y are spaces with subspaces $A \subseteq X$, $B \subseteq Y$, and that $f \,.\, X \to Y$ is a homeomorphism with $f(A) = B$. Prove that the maps $g : A \to B$ and $h : X \setminus A \to Y \setminus B$ induced by f are both homeomorphisms.

Exercise 10.11 Suppose that X, Y are topological spaces each with the discrete topology. Prove that the product topology for $X \times Y$ is discrete.

Exercise 10.12 Suppose that $\mathcal{S}$ is the Sierpinski space of Example 7.7. Find the product topology of $\mathcal{S} \times \mathcal{S}$

Exercise 10.13 Suppose that $(X, \mathcal{T}_X)$, $(Y, \mathcal{T}_Y)$ are spaces each with the co-finite topology (see Example 7.9) Show that the product topology on $X \times Y$ is not in general the co-finite topology.

[Hint: Consider $U \times Y$ in the case when U is a non-empty open set and Y is infinite.]

Exercise 10.14 The goal of this exercise is to prove Proposition 10.17. Let (X, d_X), (Y, d_Y) be metric spaces. Let $\mathcal{T}_X$ be the topology arising from d_X and $\mathcal{T}_Y$ the topology arising from d_Y. Let d denote any one of the product metrics on $X \times Y$ defined in Example 5.10, and let $\mathcal{T}_d$ be the topology on $X \times Y$ arising from d. Prove that $\mathcal{T}_d$ coincides with the product topology of the spaces $(X, \mathcal{T}_X)$, $(Y, \mathcal{T}_Y)$.

Exercise 10.15 (a) Prove that W is open in a topological product $X \times Y$ then $p_X(W)$ is open in X and $p_Y(W)$ is open in Y.

(b) Give an example of a closed set $W \subset \mathbb{R} \times \mathbb{R}$ whose projection $p_1(W)$ on the x-axis is not closed in $\mathbb{R}$.

Exercise 10.16 Suppose that X, Y are spaces and that $A \subseteq X$, $B \subseteq Y$. Prove that

(i) the interior of $A \times B$ is $\mathring{A} \times \mathring{B}$,

(ii) $\overline{A \times B} = \overline{A} \times \overline{B}$,

(iii) $\partial(A \times B) = ((\partial A) \times \overline{B}) \cup (\overline{A} \times (\partial B))$.

Exercise 10.17 Suppose that $X \times X$ is the topological product of a space X with itself. Prove that $t : X \times X \to X \times X$ defined by $t(x, x') = (x', x)$ is a homeomorphism.

Exercise 10.18 Suppose that X, X', Y, Y' are spaces and that $X \times Y$, $X' \times Y'$ have the product topologies. Suppose also that $f : X \to Y$, $g : X' \to Y'$ are homeomorphisms. Prove that if the map $f \times g : X \times Y \to X' \times Y'$ is defined by $(f \times g)(x, y) = (f(x), g(y))$ then $f \times g$ is a homeomorphism

Exercise 10.19 For each of the following maps f from a subset X of the real line into the real line, draw a rough sketch of the graph and prove that $x \mapsto (x, f(x))$ gives a homeomorphism from X onto G_f.

(a) $X = (-1, 1)$, $f(x) = \dfrac{1}{1 - x^2}$.

(b) $X = [0, \infty)$, $f(x) = \begin{cases} x \sin(1/x). & x > 0, \\ 0 & x = 0. \end{cases}$

Exercise 10.20* Prove that the topology on a space X is discrete iff the diagonal Δ is open in the topological product $X \times X$. Recall that $\Delta = \{(x, x) : x \in X\}$.

11 The Hausdorff condition

Motivation

So far in this book, we have been marching relentlessly towards generality. In this chapter we take a small step backwards, admitting that for many, though not all, purposes topological spaces are rather *too* general, and it is good to impose an extra condition on our spaces. To motivate this condition, let us try to generalize the idea of convergence of a sequence of real numbers to topological spaces.

Recall from our study of metric spaces that a sequence (x_n) in a metric space (X, d) *converges* to a point $x \in X$ if given any (real number) $\varepsilon > 0$, there exists (an integer) N such that $x_n \in B_\varepsilon(x)$ whenever $n \geqslant N$.

As we know from Chapters 6, 7, the way to generalize this to topological spaces is to replace open balls by open sets.

Definition 11.1 *A sequence of points (x_n) in a topological space X* converges *to a point $x \in X$ if given any open set $U \ni x$ there exists (an integer) N such that $x_n \in U$ whenever $n \geqslant N$.*

However, convergence in a topological space does not always satisfy our intuition, as the next example illustrates.

Example 11.2 Let X be a topological space with the indiscrete topology (see Example 7.5). Then any sequence (x_n) in X converges to any point $x \in X$. For given any open set U containing x, we must have $U = X$ (since the only open sets are $\emptyset$ and X) so $x_n \in U$ for all $n \geqslant 1$.

Let us pinpoint what has led to this nonsense, by considering how to prove uniqueness of limits of real number sequences. If we analyze the proof of Proposition 4.13, we see that it can be stated as follows. Suppose that the real number sequence (x_n) converges both to x and to y. Then if $y \neq x$, we can choose $\varepsilon = |x - y|/2$, so that $B_\varepsilon(x)$ and $B_\varepsilon(y)$ are disjoint, and since x_n is supposed to belong to each of these for sufficiently large n, we get a contradiction. In Example 11.2, on the other hand, there are no disjoint open sets U, V such that $x \in U$, $y \in V$.

Separation conditions

Definition 11.3 *A topological space X* satisfies the Hausdorff condition *if for any two distinct points $x, y \in X$ there exist disjoint open sets U, V of X such that $x \in U$, $y \in V$.*

We refer to a topological space which satisfies the Hausdorff condition as a *Hausdorff space.*

A distinguished professor used to embed this definition in his students' minds by saying that a space is Hausdorff if any two distinct points can be housed off from each other by disjoint open sets.

Proposition 11.4 *In a Hausdorff space, any given convergent sequence has a unique limit.*

Proof We simply replace the open balls $B_\varepsilon(x)$, $B_\varepsilon(y)$ in the proof at the end of the motivation section by disjoint open sets containing x, y respectively. This generalizes Proposition 4.13, the same result for sequences in $\mathbb{R}$. □

Proposition 11.5 *Any metrizable space $(X, \mathcal{T})$ is Hausdorff.*

Proof Suppose that d is a metric such that $\mathcal{T} = \mathcal{T}_d$. If $x, y \in X$ with $y \neq x$, then $d(x, y) > 0$. Take $\varepsilon = d(x, y)/2$. Then (see Exercise 5.5) the open balls $B_\varepsilon(x)$, $B_\varepsilon(y)$ are disjoint open sets containing x, y respectively. □

This of course gives us a large number of examples of Hausdorff spaces, and shows that any non-Hausdorff space is not metrizable either.

Example 11.6 Let X be any infinite space with the co-finite topology. Then X is not Hausdorff (and hence not metrizable).

Proof Suppose that x, y are distinct points of X, and let U, V be any open sets of X containing x, y respectively. Since $x \in U$, we know $U \neq \emptyset$, so since U is open in the co-finite topology we must have that $X \setminus U$ is finite. Similarly $X \setminus V$ is finite. Hence $X \setminus (U \cap V) = (X \setminus U) \cup (X \setminus V)$ is also finite. Since X is infinite, we must have $U \cap V \neq \emptyset$ (in fact it must be infinite). □

The proof of the next proposition is Exercise 11.4.

Proposition 11.7 (*a*) *Any subspace of a Hausdorff space is Hausdorff.*

(*b*) *The topological product $X \times Y$ of spaces X and Y is Hausdorff iff both X and Y are Hausdorff.*

(*c*) *If $f : X \to Y$ is an injective continuous map of topological spaces and Y is Hausdorff then so is X.*

(*d*) *If spaces X and Y are homeomorphic then X is Hausdorff iff Y is Hausdorff. In other words, Hausdorffness is a topological property.*

The Hausdorff condition is just one of a hierarchy of 'separation axioms' which a space may or may not satisfy. There is some disagreement about the names of the various conditions, including the following samples.

Definition 11.8 *A topological space is* regular, (normal) *if given any closed subset $V \subseteq X$ and point $x \in X \setminus V$ (closed set V' disjoint from V) there exist disjoint open subsets U, U' of X such that $V \subseteq U$ and $x \in U'$ ($V' \subseteq U'$).*

Exercise 11.1 Show that if X is a space with the indiscrete topology and having at least two distinct points then X is not Hausdorff.

Exercise 11.2 (a) Show that if X is a Hausdorff space, every 'singleton' set $\{x\}$ (i.e. a set containing just a single point) is closed in X.

(b) Prove that if a finite space is Hausdorff then it must have the discrete topology.

Exercise 11.3 Suppose that $x_1, x_2, \ldots, x_n$ are distinct points in a Hausdorff space X. Show that there exist pairwise disjoint open subsets $U_1. U_2, . . , U_n$ of X such that $x_i \in U_i$ for every $i \in \{1. 2, \ldots, n\}$.

Exercise 11.4 Prove Proposition 11.7.

Exercise 11.5 Suppose that $f : X \to Y$ is a continuous map of a topological space X to a Hausdorff space Y. Prove that the graph G_f of f is a closed subset of the topological product $X \times Y$.

Exercise 11.6 (a) Prove that if x is any point in a Hausdorff space X, then the intersection of all open subsets of X containing x is $\{x\}$.

(b) Give an example to show that the conclusion of (a) does not imply that X is Hausdorff.

[Hint: Think about the co-finite topology on an infinite set.]

Exercise 11.7 (a) Recall the definition $\Delta = \{(x, x) : x \in X\}$ of the diagonal subset Δ of $X \times X$. Prove that a space X is Hausdorff iff Δ is closed in the topological product $X \times X$.

(b) Let $\mathbb{S}$ be the Sierpinski space of Example 7.7. Show that a space X is Hausdorff iff the characteristic function $\chi_A : X \times X \to \mathbb{S}$ is continuous, where $A = X \times X \setminus \Delta$.

Exercise 11.8 Suppose that X, Y are spaces, with Y Hausdorff, and that A is a subspace of X. Prove that if $f, g : \overline{A} \to Y$ are continuous and $f(a) = g(a)$ for all $a \in A$ then $f = g$.

Exercise 11.9* Let X be a metric space with metric d, and for any point $x \in X$ and non-empty subset $A \subseteq X$, let $d(x, A)$ be as in Exercise 6.16 and define the map $f_A : X \to \mathbb{R}$ by $f_A(x) = d(x, A)$. Recall Exercise 6.16 asserts that f_A is continuous.

Suppose that A, B are non-empty disjoint closed sets in X. Define $g : X \to \mathbb{R}$ by $g = f_A - f_B$. Prove that $g^{-1}(-\infty, 0)$ and $g^{-1}(0, \infty)$ are disjoint open sets containing A, B respectively. (This essentially shows that any metrizable space is normal — we just have to check the special case in which at least one of A, B is empty.)

Exercise 11.10 Let $f, g : X \to Y$ be continuous maps of a topological space to a Hausdorff space. Show that $C = \{x \in X : f(x) = g(x)\}$ is closed in X. Deduce that if $f : X \to X$ is a continuous self-map of a Hausdorff space then the 'fixed-point set' $F = \{x \in X : f(x) = x\}$ is closed in X.

12 Connected spaces

Motivation

Intuitively a connected space is one which does not fall apart into two or more pieces. To make this mathematical we need a precise definition of 'fall apart'. We should probably agree that in the real line the subspace $[0, 1]$ is connected while the subspace $[0, 1] \cup [2, 3]$ is not. But what about the subspace $\mathbb{Q}$ of $\mathbb{R}$, or the subspace of the plane consisting of the graph of $y = \sin(1/x)$ for $x > 0$ together with the line segment on the y-axis joining the points $(0, -1)$ and $(0, 1)$ (see Figure 12.1)? It is not so easy to decide intuitively whether these should be considered to be connected.

In this chapter we discuss two slightly different formulations of connectedness. The first is the more basic for the study of continuity. As an application we re-prove the intermediate value theorem, Theorem 4.35. The second formulation, path-connectedness, is possibly closer to intuition. We then compare the two kinds of connectedness.

Both kinds of connectedness are used in the study of functions of a complex variable, at a few crucial stages (see Priestley 2003). In a sense they also form the beginnings of algebraic topology. For example, we shall show that connectedness of a space is a topological invariant. Thus one of the easiest ways to prove that two spaces are not homeomorphic, if it works, is to see that one is connected while the other is not.

Connectedness

From the point of view of continuity, one meaning which may be assigned to the statement that a space X 'falls apart' into two pieces A and B is this: X is the disjoint union of A and B, and moreover we can define continuous maps on X (to other spaces) whose values on A 'bear no relation' to their values on B, by which we mean that the values on A and the values on B are completely independent of each other. To be less vague, and to get down to a specific test, let $f : X \to \mathbb{R}$ be defined by:

$$f(x) = \begin{cases} 0 & \text{if } x \in A, \\ 1 & \text{if } x \in B. \end{cases}$$

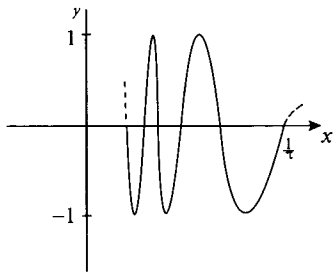

Figure 12.1. Topologist's sine curve

Then f fulfils the intuitive property that its values on A are independent of its values on B. If f is continuous on X, it is reasonable to think of A and B as mutually independent pieces of X, as far as continuity of real-valued functions on X is concerned. As we shall see in Exercise 12.2 this specific test is equivalent to any more comprehensive test along the same lines that one might envisage, of the 'independence' of A and B with regard to continuous maps on X. Moreover, in view of Proposition 10.6, it is just as good to think of our f above as taking values in the discrete subspace $\{0, 1\}$ of $\mathbb{R}$. We are now close to a definition of what it means for X not to be connected; by negating this we get the definition of connected.

Definition 12.1 *A topological space X is* connected *if there does not exist a continuous map from X* onto *a two-point discrete space.*

If you prefer to avoid negatives, you could say instead that X is connected if any continuous map from X to a two-point discrete space is constant.

Throughout the discussion of connectedness we shall give the two-point space $\{0, 1\}$ the discrete topology.

There is an equivalent of Definition 12.1 which is often taken as the definition. To state it, we need the idea of a *partition.*

Definition 12.2 *A* partition $\{A, B\}$ *of a topological space X is a pair of non-empty subsets A, B of X such that $X = A \cup B$, $A \cap B = \emptyset$, and both A and B are open in X.*

It follows that A and B are also closed in X, since they are each other's complements. The term *partition* is overworked in mathematics; for example, in set theory it would mean just a decomposition into disjoint subsets. One has to remember that here it has the meaning appropriate

to topological spaces. For this reason some prefer the term 'disconnection' to 'partition', but the latter is more usual.

Proposition 12.3 *A topological space is connected iff it admits no partition.*

Proof We prove this by showing that X is disconnected iff it does admit a partition.

First suppose $\{A, B\}$ is a partition of X. Define $f : X \to \{0, 1\}$ by

$$f(x) = \begin{cases} 0 & \text{if } x \in A, \\ 1 & \text{if } x \in B. \end{cases}$$

Since A, B are non-empty, f is onto $\{0, 1\}$. The complete list of open sets in $\{0, 1\}$ is $\{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$, and the inverse images of these under f are $\emptyset$, A, B, X which are all open in X so f is continuous. Hence by Definition 12.1 X is not connected.

Conversely suppose that $f : X \to \{0, 1\}$ is a continuous map onto the discrete two-point space. Then it is easy to see that $\{f^{-1}(0), f^{-1}(1)\}$ is a partition of X. □

Corollary 12.4 *A topological space X is connected iff the only subsets of X which are both open and closed in X are X, $\emptyset$.*

Proof We prove that X is disconnected iff there is a non-empty open and closed subset of X which is not all of X.

If $\{A, B\}$ is a partition of X, then A and B are both open and closed in X and neither is X or $\emptyset$. Conversely if A is open and closed in X and is neither $\emptyset$ nor X, then $\{A, X \setminus A\}$ is a partition of X. □

Henceforth we shall use interchangably the three forms of definition of connectedness (given in 12.1, 12.3, 12.4).

Example 12.5 (a) Any discrete space with at least two points is disconnected.

(b) Any indiscrete space is connected.

Proof (a) Such a space X admits a partition, since all its subsets are open in X.

(b) The only open sets are X and $\emptyset$, so these are the only open and closed sets. □

Definition 12.6 *A non-empty subset A of a topological space X is* connected *if A with the subspace topology is connected according to Definition 12.1. Conventionally we regard the empty set as being connected.*

Our next aim is to show that the connected subspaces of the real line (with its usual topology) are precisely the intervals listed in Chapter 2. We first characterize intervals by their property of 'betweenness'.

Proposition 12.7 *A non-empty subset $S \subseteq \mathbb{R}$ is an interval iff it satisfies the following property: if $x, y \in S$ and $z \in \mathbb{R}$ are such that $x < z < y$ then $z \in S$.*

Proof If S is one of the intervals listed in Chapter 2 then clearly it has this property. Conversely suppose that S is a non-empty subset of $\mathbb{R}$ with this property. Let

$$a = \inf S \quad \text{or } -\infty \text{ if } S \text{ is not bounded below,}$$

$$b = \sup S \quad \text{or } \infty \text{ if } S \text{ is not bounded above.}$$

We shall prove that $(a, b) \subseteq S \subseteq [a, b]$, where to avoid listing special cases we make the temporary convention that a bracket [or] placed next to $\pm\infty$ means the same as a parenthesis (or), and that if $a = b$ then (a, b) means $\emptyset$.

First, suppose $z \in (a, b)$. Then $z > a$ so by definition of a as $\inf S$ (or since S is not bounded below) there exists $x \in S$ with $x < z$. Similarly there exists $y \in S$ with $y > z$. So by hypothesis, $z \in S$. This proves $(a, b) \subseteq S$. The inclusion $S \subseteq [a, b]$ follows from the definitions of a, b. But if $(a, b) \subseteq S \subseteq [a, b]$ then S can only be one of the listed intervals. □

We can use this to prove half of the result we want.

Theorem 12.8 *Any connected subspace S of $\mathbb{R}$ is an interval.*

Proof Suppose that S is not an interval. By Proposition 12.7, there exist $x, y \in S$ and $z \in \mathbb{R} \setminus S$ with $x < z < y$. Then $\{(-\infty, z) \cap S,\ (z, \infty) \cap S\}$ is a partition of S. For by definition of the subspace topology, each of these sets is open in S; each is non-empty since they contain x, y respectively; and clearly they are disjoint and have union S, since $z \notin S$. □

Example 12.9 $\mathbb{Q}$ is not connected.

This is a special case of Theorem 12.8. An explicit partition is given by considering $\{(-\infty, \alpha) \cap \mathbb{Q},\ (\alpha, \infty) \cap \mathbb{Q}\}$ for any irrational number α.

Theorem 12.10 *Any interval I in $\mathbb{R}$ is connected.*

Proof There are various proofs of this, all ultimately based on the completeness property of real numbers.

Suppose that I is an interval in $\mathbb{R}$ and suppose for a contradiction that $\{A, B\}$ is a partition of I. Let $a \in A$, $b \in B$ and suppose without loss of generality that $a < b$. (Otherwise we may exchange the names of A and B.) Since $a, b \in I$ and I is an interval, $[a, b] \subseteq I$.

Let $A' = A \cap [a, b]$ and $B' = B \cap [a, b]$. Since A and B are closed in I and $[a, b] \subseteq I$, we may apply Exercise 10.5 to see that A' and B' are closed in $[a, b]$. Since also $[a, b]$ is closed in $\mathbb{R}$ we map apply Exercise 10.6(b) to deduce that A' and B' are closed in $\mathbb{R}$. Let $c = \sup A'$. Then $c \in A'$ since A' is closed (Example 6.8(c) and Proposition 6.11(c)). Hence $c < b$ since $b \in B'$ and $A' \cap B'$ is empty. But A' is open in $[a, b]$, so for some $\delta > 0$ we have $(c - \delta, c + \delta) \cap [a, b] \subseteq A'$. Since $c < b$, there exist points in $(c, c + \delta) \cap [a, b]$ greater than c, and such points lie in A', contradicting the choice of c. □

We may also deduce Theorem 12.10 from the intermediate value theorem (Theorem 4.35). In fact the result that intervals are connected is equivalent to the intermediate value theorem, as we shall see later.

Proof of 12.10 from the intermediate value theorem. Suppose that I is an interval and that $f : I \to \{0, 1\}$ is a continuous map onto the two-point discrete space. Let $g : I \to \mathbb{R}$ be the composition of f followed by inclusion i of $\{0, 1\}$ into $\mathbb{R}$. Since f and i are continuous, so is g. Since f is onto, $f(a) = 0$, $f(b) = 1$ for some $a, b \in I$. Now let J be the closed interval with end-points a, b and think about the intermediate value theorem applied to the real-valued continuous function g on J. Then g should take on the value $1/2$ somewhere in J. But g takes on as values only 0, 1. This contradiction shows that I is connected. □

Proposition 12.11 *Suppose that $f : X \to Y$ is a continuous map of topological spaces and that X is connected. Then $f(X)$ is connected.*

We paraphrase this result: 'a continuous image of a connected space is connected'.

Proof We first show that it is enough to prove this in the case when f is onto. For suppose f is not necessarily onto. Define $f_1 : X \to f(X)$ by $f_1(x) = f(x)$ for all $x \in X$. Then by Proposition 10.6 f_1 is continuous since f is continuous. Also, f_1 is onto. Since $f(X) = f_1(X)$, it is enough to show that $f_1(X)$ is connected. Thus in Proposition 12.11 we may replace f by f_1, and it is therefore enough to prove the proposition when f is onto.

So suppose that f is onto, suppose for a contradiction that $\{U, V\}$ is a partition of Y, and consider $\{f^{-1}(U), f^{-1}(V)\}$. Since U, V are non-empty and f is onto, $f^{-1}(U)$, $f^{-1}(V)$ are also non-empty. Since f is continuous and U, V are open in Y, it follows that $f^{-1}(U)$, $f^{-1}(V)$ are open in X. Since $U \cap V = \emptyset$. also $f^{-1}(U) \cap f^{-1}(V) = f^{-1}(U \cap V) = \emptyset$. Since $U \cup V = Y$, also $f^{-1}(U) \cup f^{-1}(V) = f^{-1}(U \cup V) = f^{-1}(Y) = X$. But this means that $\{f^{-1}(U), f^{-1}(V)\}$ is a partition of X, and this contradiction proves the result. □

Corollary 12.12 *Connectedness is a topological property.*

Proof Suppose that $f : X \to Y$ is a homeomorphism of spaces. If X is connected then so is Y by Proposition 12.11. Similarly if Y is connected then X is connected by Proposition 12.11 since $f^{-1} : Y \to X$ is continuous and onto. □

Corollary 12.13 *With the hypotheses of 12.11 the graph $\mathcal{G}_f$ of f is connected.*

Proof This follows from 12.12 since $\mathcal{G}_f$ is homeomorphic to X by Proposition 10.18. □

Corollary 12.14 *Suppose that $f : X \to \mathbb{R}$ is continuous and X is connected. Then $f(X)$ is an interval.*

Proof This follows from Proposition 12.11 since connected subsets of $\mathbb{R}$ are intervals. □

Corollary 12.15 (Intermediate value theorem) *If $f : [a, b] \to \mathbb{R}$ is continuous then it has the intermediate value property.*

Proof This follows from the previous corollary since intervals have the betweenness property. □

Next we prove some general results about connectedness.

Proposition 12.16 *Suppose that $\{A_i : i \in I\}$ is an indexed family of connected subsets of a topological space X with $A_i \cap A_j \neq \emptyset$ for each pair $i, j \in I$. Then $\bigcup_{i \in I} A_i$ is connected.*

Proof Suppose that $f : \bigcup_{i \in I} A_i \to \{0, 1\}$ is continuous. For each $i \in I$ the restriction $f|_{A_i} : A_i \to \{0, 1\}$ is continuous, hence constant since A_i is connected. Moreover, for any $i, j \in I$ we have $A_i \cap A_j \neq \emptyset$, so the constant

values f that takes on A_i, A_j are the same. Hence f is constant, so $\bigcup_{i\in I} A_i$ is connected. □

Corollary 12.17 *Suppose that $\{C_i : i \in I\}$ and B are connected subsets of a space X such that for every $i \in I$ we have $C_i \cap B \neq \emptyset$. Then $B \cup \bigcup_{i\in I} C_i$ is connected.*

Proof First apply 12.16 to the pair B, C_i for a particular $i \in I$, to get that $B \cup C_i$ is connected. Then apply 12.16 to the family $\{B \cup C_i : i \in I\}$. □

Theorem 12.18 *The topological product $X \times Y$ of spaces X, Y is connected iff both X, Y are connected.*

Proof Recall in this proof that by definition X and Y are non-empty.

First, if $X \times Y$ is connected, so are X and Y since these are the continuous images of $X \times Y$ under the projections p_X, p_Y (see Proposition 10.10).

Conversely suppose that X and Y are connected. For each $y \in Y$ the subset $X \times \{y\}$ of $X \times Y$ is connected by 12.11 since it is the continuous image of $i_y : X \to X \times \{y\}$ given by $i_y(x) = (x, y)$. Similarly, for fixed $x_0 \in X$, the subset $\{x_0\} \times Y$ of $X \times Y$ is connected. But for any y the intersection $X \times \{y\} \cap \{x_0\} \times Y = \{(x_0, y)\}$ is non-empty, and the result follows from Corollary 12.17 since $X \times Y = (\{x_0\} \times Y) \cup \bigcup_{y\in Y} X \times \{y\}$. □

Proposition 12.19 *Suppose that A is a connected subset of a space X and that $A \subseteq B \subseteq \overline{A}$. Then B is connected.*

Proof Suppose that $f : B \to \{0, 1\}$ is continuous. Then $f|_A$ is constant since A is connected. Suppose without loss of generality that $f(a) = 0$ for all $a \in A$. (Otherwise interchange the roles of 0 and 1.) Now suppose for a contradiction that $f(b) = 1$ for some $b \in B$. Since $\{1\}$ is open in $\{0, 1\}$, $f^{-1}(1)$ is open in B so $f^{-1}(1) = B \cap U$ for some U open in X. Now U is open and contains the point b which is in $\overline{A}$, so $U \cap A \neq \emptyset$, say $a \in U \cap A \subseteq U \cap B = f^{-1}(1)$. So $f(a) = 1$, which is a contradiction. Hence B is connected. □

Path-connectedness

There is another kind of connectedness which is probably more intuitive than the kind we have looked at so far.

Definition 12.20 *For points x, y in a topological space X, a* path in X from x to y *is a continuous map $f : [0, 1] \to X$ such that $f(0) = x$, $f(1) = y$. We say that such a path* joins x and y.

Definition 12.21 *A topological space X is* path-connected *if any two points of X can be joined by a path in X.*

As in the case of 'connected', we say that a non-empty subset $A \subseteq X$ is path-connected if with the subspace topology it satisfies 12.21, and conventionally we say the empty set is path-connected.

Example 12.22 (a) For any $n \geqslant 1$, $\mathbb{R}^n$ is path-connected. More generally any convex subset C of $\mathbb{R}^n$ is path-connected.

(b) Any annulus in $\mathbb{R}^2$ is path-connected, where an annulus means a ring-shaped set of the form $\{(x, y) \in \mathbb{R}^2 : a \leqslant (x-c)^2 + (y-d)^2 \leqslant b\}$ for some real numbers a, b, c, d with $0 < a < b$).

Proof (a) By definition of convexity, any two points in C may be joined by a straight line segment in C.

(b) This is Exercise 12.14. □

Several of the results we have proved for connectedness are also true for path-connectedness, and these are on the web site.

Comparison of definitions

Proposition 12.23 *Any path-connected space X is connected.*

Proof This follows from connectedness of the unit interval in the real line. For suppose X is path-connected and $g : X \to \{0, 1\}$ is continuous. Suppose for a contradiction that g is not constant, so there exist x, $y \in X$ with $g(x) = 0$ and $g(y) = 1$. Let $f : [0, 1] \to X$ be a path in X from x to y. Then the composition $g \circ f : [0, 1] \to \{0, 1\}$ is continuous onto, contradicting connectedness of $[0, 1]$. □

This proposition gives the main way of checking that a space is connected- we show that it is path-connected hence connected. This is not infallible, since a space can be connected without being path-connected (this is true of the space indicated in Figure 12.1, as explained on the web site).

Next here is a useful result which is particular to path-connectedness; it states the intuitively obvious fact that if we go along a path from x to y and then along a path from y to z, the result is a path from x to z.

Lemma 12.24 *Suppose that $f, g : [0, 1] \to X$ are paths in a space X from x to y and from y to z, respectively. Let*

$$h(x) = \begin{cases} f(2t) & t \in [0, 1/2], \\ g(2t-1) & t \in [1/2, 1]. \end{cases}$$

Then h is a path in X from x to z.

Proof First, h is well defined, since when $t = 1/2$ the two parts of the definition of h 'fit together': $f(2t) = f(1) = y$, $g(2t-1) = g(0) = y$. Also, the restriction $h|[0, 1/2] = f \circ k$ where $k : [0, 1/2] \to [0, 1]$ is given by $k(t) = 2t$. Since f and k are both continuous, $h|[0, 1/2]$ is continuous. Likewise $h|[1/2, 1]$ is continuous. Now h is continuous by Exercise 10.7(b). Finally, h is a path in X from x to z, since we have $h(0) = f(0) = x$ and $h(1) = g(1) = z$. □

Although the converse of Proposition 12.23 is false in general, we do have the following result.

Proposition 12.25 *A connected open subset U of $\mathbb{R}^n$ is path-connected.*

Proof Figure 12.2 may help in following this proof.

If U is empty, the result is true by convention. Otherwise choose a point $x_0 \in U$. We shall show that the set $V \subseteq U$ of all points in U which can be joined to x_0 by a path in U, and its complement $U \setminus V$ are both open. Since they are clearly disjoint and have union U, they would form a partition of U if they were both non-empty. So one of them must be empty. But $x_0 \in V$, so $V = U$, and this will give the result.

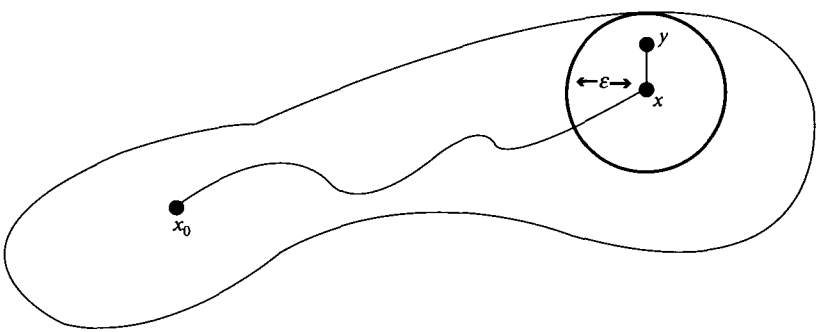

Figure 12.2. A connected open subset of the plane is path-connected

To show that V is open in U (in fact in $\mathbb{R}^n$) suppose $x \in V$. Then $x \in U$ and since U is open in $\mathbb{R}^n$, there is some $\varepsilon > 0$ such that $B_\varepsilon(x) \subseteq U$. Now any point y in $B_\varepsilon(x)$ may be joined to x by a straight line segment in $B_\varepsilon(x)$ and hence in U, so using Lemma 12.24 we see that any point in $B_\varepsilon(x)$ may be joined by a path in U to x_0. Hence $B_\varepsilon(x) \subseteq V$ and we see that V is open in $\mathbb{R}^n$ and hence in U.

We show that $U \setminus V$ is also open in U by a similar argument. Suppose that $x \in U \setminus V$, and let $\varepsilon > 0$ be such that $B_\varepsilon(x) \subseteq U$. If any point y in $B_\varepsilon(x)$ could be joined to x_0 by a path in U, then so could x by composing with the straight line segment from y to x. So $B_\varepsilon(x) \subseteq U \setminus V$, and $U \setminus V$ is open in $\mathbb{R}^n$ and hence in U. By the commentary above, this completes the proof. □

The same proof shows that any two points in a connected open subset U of $\mathbb{R}^n$ may be joined by a polygonal path (the juxtaposition of finitely many straight line segments) in U, or even one in which each segment is parallel to one of the coordinate axes.

► Proposition 12.25 remains true, with essentially the same proof, when $\mathbb{R}^n$ is replaced by any normed vector space. ◄

We have looked at when a space intuitively falls apart or is all of one piece. More generally we can consider 'how many pieces' it falls into; this leads to the idea of *components*, discussed on the web site.

Connectedness and homeomorphisms

It is intuitively clear that $\mathbb{R}$ and $\mathbb{R}^2$ are not homeomorphic. Here is one way to prove this: suppose for a contradiction that $f : \mathbb{R} \to \mathbb{R}^2$ is a homeomorphism. Then by Exercise 10.10, f induces a homeomorphism $f|\mathbb{R} \setminus \{0\} : \mathbb{R} \setminus \{0\} \to \mathbb{R}^2 \setminus \{f(0)\}$. But $\mathbb{R} \setminus \{0\}$ is not connected whereas $\mathbb{R}^2$ with a point removed is path-connected and hence connected. This is a contradiction by Corollary 12.12, so $\mathbb{R}$ and $\mathbb{R}^2$ are not homeomorphic.

We can argue similarly in many cases. For another example, the interval $I = [0, 1]$ and the circle $S^1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ are not homeomorphic. For suppose there were a homeomorphism $f : I \to S^1$. Then there would be an induced homeomorphism of $[0, 1/2) \cup (1/2, 1]$ to $S^1 \setminus \{f(1/2)\}$. But the first of these two spaces is disconnected whereas S^1 with a point removed is path-connected hence connected. The web site has further examples of this line of argument.

Exercise 12.1 Which of the following subsets of $\mathbb{R}^2$ are (a) path-connected (b) connected?

(i) $\overline{B_1((1,0))} \cup B_1((-1,0))$;
(ii) $\overline{B_1((1,0))} \cup \overline{B_1((-1,0))}$;
(iii) $B_1((1,0)) \cup B_1((-1,0))$;
(iv) the 'rational comb' $\{(q, y) \in \mathbb{R}^2 \,.\, q \in \mathbb{Q}, y \in [0,1]\} \cup (\mathbb{R} \times \{1\})$;
(v) the set of all points in $\mathbb{R}^2$ with at least one coordinate in $\mathbb{Q}$.

Exercise 12.2 Suppose that $\{A, B\}$ is a partition of a topological space X and that $f : X \to Y$ is a map to another space Y. Prove that if the restrictions $f|A$ and $f|B$ are both continuous then f is continuous.

Exercise 12.3 Prove that any infinite set with the co-finite topology (see Example 7.9) is connected.

Exercise 12.4 Suppose that $\mathcal{T}_1, \mathcal{T}_2$ are two topologies for a set X, and that $\mathcal{T}_1 \subseteq \mathcal{T}_2$. Does it follow that $(X, \mathcal{T}_1)$ is connected if $(X, \mathcal{T}_2)$ is connected? Does it follow that $(X, \mathcal{T}_2)$ is connected if $(X, \mathcal{T}_1)$ is connected?

Exercise 12.5 Suppose that for each $i \in \{1, 2, \ldots, n\}$ that A_i is a connected subset of a space X, such that $A_i \cap A_{i+1} \neq \emptyset$ for each $i \in \{1, 2, \ldots, n-1\}$. Prove that $\bigcup_{i=1}^{n} A_i$ is connected. Does this result extend to an infinite sequence (A_i) of connected subsets?

Exercise 12.6 A map $X \to \mathbb{R}$ from a space X is said to be *locally constant* if for each $x \in X$ there is some open set U with $x \in U$ and $f|U$ constant. Prove that if X is connected then every locally constant map $f : X \to \mathbb{R}$ is constant. Is this still true if $\mathbb{R}$ is replaced by an arbitrary topological space?

The next few exercises, 12.7-12.10, can be done using just the intermediate value theorem; as we have seen this is intimately related to connectedness of real intervals.

Exercise 12.7 Show that any polynomial equation of odd degree with real coefficients has at least one real root.

Exercise 12.8 Prove that any continuous function $f : [a, b] \to [a, b]$ has a fixed point, that is a point $x \in [a, b]$ such that $f(x) = x$.

Exercise 12.9 Suppose that $f : [0, 1] \to \mathbb{R}$ is continuous and that $f(1) = f(0)$. Show that for $n \in \mathbb{N}$, $n \geqslant 2$ there exists $x \in [0, 1]$ such that $f(x + 1/n) = f(x)$. [Hint: For $g(x) = f(x) - f(x + 1/n)$, show $g(0) + g(1/n) + \ldots + g((n-1)/n) = 0$. Now observe that either all the $g(i/n)$ are zero, or two successive ones have opposite signs.]

Exercise 12.10* Suppose that $f : \mathbb{R} \to \mathbb{R}$ is a function such that for every $x \in \mathbb{R}$ the set $f^{-1}(x)$ contains exactly two points. Show that f cannot be continuous.

Exercise 12.11 Give either a proof of, or a counterexample to, each of the following.
(a) Suppose that X, Y are spaces with subsets A, B. Suppose that neither $X \setminus A$ nor $Y \setminus B$ is connected. Then $X \times Y \setminus (A \times B)$ is not connected.
(b) Suppose that A, B are subsets of a space X and that both $A \cap B$ and $A \cup B$ are connected. Then A and B are connected.
(c) Suppose that A, B are closed subsets of a space X and that both $A \cap B$ and $A \cup B$ are connected. Then A and B are connected.

Exercise 12.12 Prove that the function space $C[0, 1]$ of all continuous real-valued functions on $[0, 1]$ with the sup metric (see Example 5.13) is path-connected and hence connected.

Exercise 12.13 Suppose that $f : X \to Y$ is a continuous map from a path-connected space X onto a space Y. Show that Y is path-connected.

Exercise 12.14 Prove Example 12.22(b), that an annulus in $\mathbb{R}^2$, (i.e. a set of the form $\{(x, y) \in \mathbb{R}^2 : a \leqslant (x - c)^2 + (y - d)^2 \leqslant b\}$ for some real numbers a, b, c, d with $0 < a < b$) is path-connected.

Exercise 12.15 Prove that a space X is connected iff every non-empty proper subset of X has non-empty boundary.
[Hint: See Exercise 9.14(b).]

Exercise 12.16 Suppose that A, B are subsets of a space X with B connected and suppose that B has a non-empty intersection with each of A and $X \setminus A$. Prove that B must also have non-empty intersection with ∂A.
[Hint: Recall Exercises 9.15 and 9.16.]

Exercise 12.17 Suppose A, B are connected subsets of a space X such that $A \cap \overline{B} \neq \emptyset$. Prove that $A \cup B$ is connected.

Exercise 12.18 Given any space X, prove that there is a connected space Y containing X as a subspace and such that $Y \setminus X$ consists of a single point.
[Hint: Let $t \notin X$ and consider the family of subsets of $X \cup \{t\}$ consisting of $\emptyset$ together with $\{U \cup \{t\} : U \in \mathcal{T}\}$ where $\mathcal{T}$ is the topology on X.]

Exercise 12.19* Give an example of a sequence of closed connected subsets V_n of the Euclidean plane such that $V_n \supseteq V_{n+1}$ for each $n \in \mathbb{N}$ but $\bigcap_{n=1}^{\infty} V_n$ is not connected.

13 Compact spaces

Motivation

The subject matter of this chapter is probably the most important single topic in this book. There is more than one way of framing the definition of compactness. The definition in this chapter is appropriate for topological spaces. Another important definition, which works well in metric spaces, will be studied in Chapter 14 and related to the present definition.

Recall from the introduction that we are aiming to prove some basic results about continuous functions in a general setting, and that the following is an example of the kind of result we wish to generalize.

Proposition 13.1 *A continuous function $f : [a, b] \to \mathbb{R}$ is bounded on $[a, b]$.*

We shall begin with a slow build-up towards one way of proving this and generalizations of it. Let us set out by supposing that the function $f : A \to \mathbb{R}$ is defined on some general subset A of $\mathbb{R}$. We ask 'Is f bounded on A?', in other words 'Does there exist a fixed real number K such that $|f(x)| \leqslant K$ for all $x \in A$?' We go from the known to the unknown in tackling this question.

STEP 1. If A is a finite set, say that $A = \{a_1, a_2, \ldots, a_r\}$ then the answer is 'Yes'. We may take $K = \max\{|f(a_1)|, |f(a_2)|, \ldots, |f(a_r)|\}$.

STEP 2. If A is a finite union of subsets, $A = \bigcup_{i=1}^{r} A_i$ and if we know that $f|A_i$ is bounded for each $i \in \{1, 2, \ldots, r\}$, say $|f(x)| \leqslant K_i$ for all $x \in A_i$, then again the answer is 'Yes'. We may take $K = \max\{K_1, K_2, \ldots, , K_r\}$. Then any x in A is in A_i for some $i \in \{1, 2, \ldots, r\}$, and $|f(x)| \leqslant K_i \leqslant K$.

Example 13.2 We now look at an example of a function which is not bounded, although it is continuous. Let $A = (0, 1)$ and define $f(x) = 1/x$ for $x \in A$. Given any real number K we can find a real number $x \in (0, 1)$ such that $0 < x < 1/K$, so $f(x) = 1/x > K$. So no K is large enough to bound f on all of $(0, 1)$.

STEP 3. However, we do get *something* when f is continuous. For suppose that $A \subseteq \mathbb{R}$ and $f : A \to \mathbb{R}$ is continuous. Let us apply the $\varepsilon - \delta$ definition of continuity at a point $a \in A$, with $\epsilon = 1$ say: thus there exists $\delta > 0$ such that $|f(x) - f(a)| < 1$ whenever $x \in A$ and $x \in B_\delta(a)$. Hence for all such x we have $|f(x)| = |f(x) - f(a) + f(a)| \leqslant |f(x) - f(a)| + |f(a)| < 1 + |f(a)|$. The δ here in general depends on a (and also on f). Let us write it in the meantime as $\delta(a)$. So for continuous f, given any $a \in A$ there exists a single bound $K_a = 1 + |f(a)|$ for $|f(x)|$ which works for all x in some neighbourhood of a (precisely, on $A \cap B_{\delta(a)}(a)$).

STEP 4. Now recall that the original question is whether there is a single bound for $|f(x)|$ which serves for all $x \in A$. We cannot in general answer this affirmatively by taking the maximum of the K_a in Step 3, because in general there are infinitely many K_a involved (one for each $a \in A$) and the set of these K_a may not be bounded above. However, suppose for a moment that A is contained in the union of some *finite* number of the $B_{\delta(a)}(a)$ occurring in Step 3. Then, since f is bounded on each $A \cap B_{\delta(a)}(a)$, it follows by Step 2 that f is bounded on A.

Let us examine more closely the assumption that allowed us to reach this conclusion: originally we just had $A \subseteq \bigcup_{a \in A} B_{\delta(a)}(a)$, and we then assumed that A is contained in the union of just *finitely many* of these neighbourhoods. Now $B_{\delta(a)}(a)$ depended on f as well as a, so if we want to set down a condition *on* A which will enable us to prove by the above argument that *any* continuous function of A is bounded, we had better assume something like the following property.

PROVISIONAL DEFINITION A subset $A \subseteq \mathbb{R}$ is *compact* if whenever it is contained in the union of a family of open balls, it is contained in the union of finitely many of these balls.

The above discussion shows that if $A \subseteq \mathbb{R}$ is compact in the above sense and $f : A \to \mathbb{R}$ is continuous then f is bounded on A.

As it stands, our provisional definition makes sense with $\mathbb{R}$ replaced by any metric space. As usual, replacing open balls by open sets generalizes it to any topological space.

Before stating the definition precisely, let us consider what compactness enables us to do. The conclusion of Step 3 is called a 'local' statement, because it asserts something only for a neighbourhood of each point. On the other hand, the statement that f is bounded on A is called a 'global' statement, since it describes a property of f on the whole domain A. Compactness of A allows us to pass from the local to the global, in dealing with continuous functions.

Definition of compactness

The definition of compactness may conveniently be expressed in the language of covers.

Definition 13.3 *Suppose X is a set and $A \subseteq X$. A family $\{U_i : i \in I\}$ of subsets of X is called a* cover for A *if $A \subseteq \bigcup_{i \in I} U_i$.*

Definition 13.4 *With notation as in Definition 13.3, a* subcover *of a cover $\{U_i : i \in I\}$ for A is a subfamily $\{U_j : j \in J\}$ for some subset $J \subseteq I$ such that $\{U_j : j \in J\}$ is still a cover for A. We call it a* finite subcover *if J is finite.*

Definition 13.5 *If $\mathcal{U} = \{U_i : i \in I\}$ is a cover for a subset A of a topological space X and if each U_i is open in X then $\mathcal{U}$ is called* an open cover *for A.*

Definition 13.6 *A subset A of a topological space X is* compact *if every open cover for A has a finite subcover.*

In particular, 13.6 gives a definition of a space X being compact. This rather intricate definition takes a bit of getting used to. It is important to notice precisely what it is saying: given *any* open cover $\mathcal{U}$ of A, there is a finite subfamily of $\mathcal{U}$ which is enough to cover A. For example, this is very different from saying that 'A has a finite open cover' indeed, the latter is true for any subset A of a space X, since $\{X\}$ is such a finite open cover. To emphasize that you have to allow yourself to set out from *any* open cover, here is an example of a non-compact subset of $\mathbb{R}$.

Example 13.7 The open interval $(0, 1)$ in $\mathbb{R}$ (with its usual topology) is not compact.

Proof Notice that if we begin with the open cover $\{(0, 1)\}$ then of course it has a finite subcover—it is itself finite. We deliberately look for an 'awkward' open cover with no finite subcover. Consider for example the family of open sets $\{(1/n, 1) : n \in \mathbb{N}, n > 1\}$. This does cover $(0, 1)$: given any $x \in (0, 1)$ we have $x > 1/n$ for sufficiently large n, and then $x \in (1/n, 1)$. But any finite subfamily, say $\{(1/n_1, 1), (1/n_2, 1), \ldots, (1/n_r, 1)\}$ covers only $(1/N, 1)$ where $N = \max\{n_1, n_2, \ldots, n_r\}$. So $(0, 1)$ is *not* compact. □

On the other hand, an important theorem which we shall prove shortly says that any closed bounded interval $[a, b]$ in $\mathbb{R}$ is compact. In the meantime, we give a couple of easier examples.

Example 13.8 (a) Any finite subset $A = \{a_1, a_2, \ldots, a_n\}$ of a space X is compact.

(b) Any space with the co-finite topology is compact.

Proof (a) Suppose $\mathcal{U} = \{U_i : i \in I\}$ is any open cover of A. Then for each $r \in \{1, 2, \ldots, n\}$ we have $a_r \in U_{i_r}$ for some $i_r \in I$, and then $\{U_{i_1}, U_{i_2}, \ldots, U_{i_n}\}$ is a finite subcover of $\mathcal{U}$ for A.

(b) Suppose X is a space with the co-finite topology, and let $\mathcal{U}$ be any open cover of X. We know that at least one of the sets in $\mathcal{U}$, say U_{i_0}, is non-empty, since X is non-empty. Since U_{i_0} is open, $X \setminus U_{i_0}$ must be finite, say $X \setminus U_{i_0} = \{x_1, x_2, \ldots, x_n\}$. For each $r = 1, 2, \ldots, n$ we have $x_r \in U_{i_r}$ for some $i_r \in I$. Then $\{U_{i_0}, U_{i_1}, \ldots, U_{i_n}\}$ is a finite subcover of $\mathcal{U}$ for X. □

Now that we have a definition of compactness, we may ask 'What is it good for?' There are at least two general answers worth thinking about.

(1) It allows us to pass from the local to the global in the sense explained in the section on motivation. In particular, the discussion there essentially proved that any real-valued continuous function on a compact space is bounded, though we shall shortly repeat that proof.

(2) The second answer has been well expressed in Hewitt (1960): compactness is the next best thing to finiteness, as far as continuous functions are concerned. Hewitt points out that many statements about functions $f : X \to Y$ are:
   (i) true and trivial if X is a finite set;
   (ii) true for continuous f when X is compact;
   (iii) false, or very hard to prove, even for continuous f, when X is non-compact.

We have already seen (2) illustrated for the statement 'f is bounded on X' when $Y = \mathbb{R}$.

► In a sense similar to (2), finite-dimensionality of a vector space is a substitute for finiteness when we are dealing with linear transformations. Consider for example the result that a linear transformation from a finite-dimensional vector space to itself is injective iff it is onto. ◄

Remarks (a) Some textbooks take 'compact' to mean our Definition 13.6 plus the Hausdorff condition.

(b) The mathematician H. Weyl is credited with a striking comment about compact subsets of the plane: 'If a city is compact, it can be guarded by a finite number of arbitrarily short-sighted policemen'.

Compactness of closed bounded intervals

Theorem 13.9 *Any closed bounded interval* $[a, b]$ *in* $\mathbb{R}$ *is compact.*

Here is a proof which uses the completeness property of $\mathbb{R}$ rather directly. It is sometimes called 'the creeping method'. There is a further proof in Exercise 13.15.

Proof Suppose that $\mathcal{U}$ is a cover of $[a, b]$ by sets open in $\mathbb{R}$. Let

$$G = \{x \in \mathbb{R} : x \geqslant a \text{ and } [a, x] \text{ is covered by a finite subfamily of } \mathcal{U}\}.$$

Let us call points in G *good (for* $\mathcal{U}$*)*. We want to show that b is good. We note that if x is good and $a \leqslant y \leqslant x$ then y is also good: for $[a, y]$ is covered by any finite subfamily of $\mathcal{U}$ that covers $[a, x]$. Since [a, a] is contained in a single set of $\mathcal{U}$, we know that $a \in G$ so G is non-empty.

If G is not bounded above, then there exists $x \in G$ with $x > b$, so b is good.

Suppose that G is bounded above, and let c be $\sup G$. If $c > b$ then we may choose $x \in G$ with $x > b$, and again b is good.

Suppose for a contradiction that $c \leqslant b$. Note that $c \neq a$ since $a \in U_a$ for some $U_a \in \mathcal{U}$, and since U_a is open there exists $\delta > 0$ with $[a, a+\delta) \subseteq U_a$, so all points in $[a, \delta)$ are good. Hence $c \in (a, b]$.

Now there is some $U_c \in \mathcal{U}$ with $c \in U_c$. Since U_c is open, there is some $\varepsilon > 0$ such that $(c - \varepsilon, c + \varepsilon) \subseteq U_c$. Since $c > a$, and by definition of $c = \sup G$, there exists $x \in G$ such that $x > a$ and $x > c - \varepsilon$. Since x is good, $[a, x]$ is covered by a finite subfamily of $\mathcal{U}$; if we add U_c to this finite subfamily we get a finite subfamily of $\mathcal{U}$ which covers $[a, c + \varepsilon/2]$. So $c + \varepsilon/2$ is good, contradicting the fact that c is an upper bound for G. Hence $c > b$ and as we have seen this shows that b is good as required. □

You may find it instructive to examine where this proof breaks down if you try to apply it to $(a, b]$ or $[a, b)$.

Properties of compact spaces

Which subsets C of $\mathbb{R}^n$ are compact? In this section, we show that it is necessary for C to be bounded, and closed in $\mathbb{R}^n$. (These are corollaries of more general results.) Later we show that for subspaces of $\mathbb{R}^n$ these conditions are also sufficient. However, as we shall see they are not sufficient in general metric spaces although they make sense there.

Proposition 13.10 *Any compact subset* C *of a metric space* (X, d) *is bounded.*

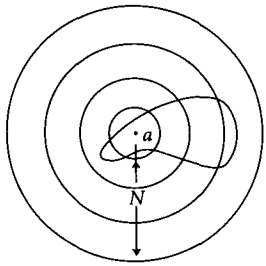

Figure 13.1. Compact subsets of metric spaces are bounded

Proof Let a be any point in X. For any $c \in C$ we may choose an integer $n > d(a, c)$ and then $c \in B_n(a)$. This shows that $\{B_n(a) : n \in \mathbb{N}\}$ is an open cover of C. By compactness, there is some finite subcover of C, say $\{B_{n_1}(a), B_{n_2}(a), \ldots, B_{n_r}(a)\}$. But $\bigcup_{i=1}^{r} B_{n_i}(a) = B_N(a)$ where N is the maximum of $\{n_1, n_2, \ldots, n_r\}$, so $C \subseteq B_N(a)$ and C is bounded as required (Figure 13.1). □

Corollary 13.11 *Any compact subset of* $\mathbb{R}^n$ *is bounded.*

Proposition 13.12 *Let C be a compact subset of a Hausdorff space X. Then C is closed in X.*

Proof Let x be some fixed point in $X \setminus C$. We shall show that there exists an open set U_x containing x and with $U_x \subseteq X \setminus C$; then by Proposition 7.2 $X \setminus C$ is open, and hence C is closed, in X.

For each $c \in C$, by the Hausdorff condition there exist disjoint sets U_c, V_c open in X and with $x \in U_c$, $c \in V_c$.

At this stage we note that if C were *finite*, say $C = \{c_1, c_2, \ldots, c_r\}$ we could take $U_x = \bigcap_{i=1}^{r} U_{c_i}$ and achieve our goal of getting an open set U_x with $x \in U_x$ and $U_x \cap C = \emptyset$. Figure 13.2 illustrates this in the case when $r = 2$.

However, more generally we can use compactness as a substitute for finiteness. For $\{V_c : c \in C\}$ is an open cover of C, and by compactness there is a finite subcover, say $\{V_{c_1}, V_{c_2}, \ldots, V_{c_r}\}$. Let $U_x = \bigcap_{i=1}^{r} U_{c_i}$. As a finite intersection of open sets, U_x is open in X. Clearly $x \in U_x$, since $x \in U_{c_i}$ for all i. We shall see that $U_x \subseteq X \setminus C$.

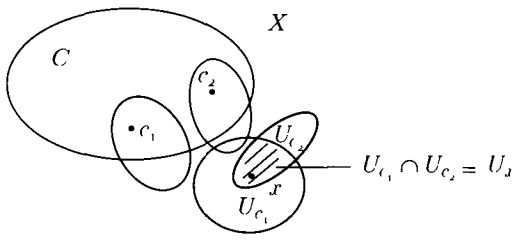

Figure 13.2. Proof of Proposition 13.12

For each $i = 1, 2, \ldots, r$, we have $U_x \subseteq U_{c_i}$, so $U_x \cap V_{c_i} \subseteq U_{c_i} \cap V_{c_i} = \emptyset$. Hence

$$U_x \cap C \subseteq U_x \cap \left(\bigcup_{i=1}^{r} V_{c_i} \right) = \bigcup_{i=1}^{r} (U_x \cap V_{c_i}) = \emptyset, \quad \text{so } U_x \subseteq X \backslash C \text{ as required.}$$

□

Corollary 13.13 *Any compact subset of* $\mathbb{R}^n$ *is closed in* $\mathbb{R}^n$.

Proposition 13.12 gives another proof that singleton point sets (and more generally finite sets) in a Hausdorff space are closed.

There is one more definition related to compactness which is sometimes useful.

Definition 13.14 *A subset A of a topological space X is said to be* relatively compact *in X if* $\overline{A}$ *is compact, where the closure is taken in X.*

For example, (0, 1) is relatively compact in $\mathbb{R}$ since [0, 1] is compact, but it is not relatively compact in (0, 1) since there its closure is still just (0, 1).

Continuous maps on compact spaces

Proposition 13.15 *If* $f : X \to Y$ *is a continuous map of topological spaces and X is compact then* $f(X)$ *is compact.*

Proof Suppose that $\mathcal{U}$ is an open cover of $f(X)$. Since f is continuous, $f^{-1}(U)$ is open in X for every $U \in \mathcal{U}$. The family $\{f^{-1}(U) : U \in \mathcal{U}\}$ covers X since $\mathcal{U}$ covers $f(X)$. Hence by compactness of X, there is a finite subcover, say $\{f^{-1}(U_1), f^{-1}(U_2), \ldots, f^{-1}(U_r)\}$ and then $\{U_1, U_2, \ldots, U_r\}$ is a finite subcover of $f(X)$. □

This result is often stated as 'the continuous image of a compact space is compact'.

Corollary 13.16 *Compactness is a topological property.*

This follows just as the analogous result for connectedness did.

Corollary 13.17 *Any continuous map from a compact space to a metric space is bounded.*

This follows since the image is compact and hence bounded. It proves Proposition 13.1 in a general setting.

Next we consider a bounded real-valued function $f : X \to \mathbb{R}$ on a space X. Since X is non-empty so too is $f(X)$; since also $f(X)$ is bounded, $\sup f(X)$ and $\inf f(X)$ exist. These are called 'the bounds of f on X' and in general they may or may not be in the set $f(X)$. If they are, we say 'f attains its bounds on X'.

Corollary 13.18 *If $f : C \to \mathbb{R}$ is continuous and C is compact then f attains its bounds on C. This means there is at least one $c_0 \in C$ such that $f(c_0) = \inf f(C)$ and at least one $c_1 \in C$ such that $f(c_1) = \sup f(C)$.*

Proof From Proposition 13.15 we know that $f(C)$ is compact, and hence bounded and closed in $\mathbb{R}$ by Corollary 13.11 and Corollary 13.13. But for any non-empty bounded subset $A \subseteq \mathbb{R}$ we know that $\sup A, \inf A \in \overline{A}$ (see Exercise 6.9) so here $\sup f(C)$ and $\inf f(C)$ are in $\overline{f(C)} = f(C)$, in other words, f attains its bounds on C. □

Corollary 13.19 *A continuous real-valued function on $[a, b]$ attains its bounds.*

Compactness of subspaces and products

We saw above that any compact subset of $\mathbb{R}^n$ is bounded and closed in $\mathbb{R}^n$. In the next section, we show conversely that closed bounded subsets of $\mathbb{R}^n$ are compact. In order to do this, it is convenient to prove two general results.

Proposition 13.20 *Any closed subset C of a compact space X is compact.*

Proof Let $\mathcal{U}$ be any cover of C by sets open in X. Since C is closed in X, $X \setminus C$ is open in X. If we add it to $\mathcal{U}$ we get an open cover of X. But X is compact, so there is a finite subcover, say $\{U_1, U_2, \ldots, U_r\}$. This certainly covers C since it covers all of X. Moreover, if $X \setminus C$ is one

of these U_i then we may throw it out and the remaining $r-1$ sets will still cover C. If $X \setminus C$ is not one of the U_i then we leave $\{U_1, U_2, \ldots, U_r\}$ alone. In either case we get a finite subcover of $\mathcal{U}$ for C. So C is compact. □

Theorem 13.21 *A topological product $X \times Y$ of spaces X, Y is compact iff both X and Y are compact.*

Proof In one direction the proof is easy: if $X \times Y$ is compact then X and Y are compact as the continuous images of $X \times Y$ under the projection maps p_X, p_Y.

Now suppose that X and Y are compact. The proof that $X \times Y$ is compact is one of the trickier proofs in topology at this level. We shall take it in easy stages.

Let $\mathcal{W}$ be any open cover of $X \times Y$. We shall call a subset $A \subseteq X$ *good* (*for* $\mathcal{W}$) if $A \times Y$ is covered by a finite subfamily of $\mathcal{W}$. We want to prove that X is good.

STEP 1. If $A_1, A_2, \ldots, A_r \subset X$ are all good, then so is their union A. For given any $i \in \{1, 2, \ldots, r\}$ there is a finite subfamily say $\mathcal{W}_i$ of $\mathcal{W}$ which covers $A_i \times Y$. Then $A \times Y$ is covered by the finite subfamily $\mathcal{W}_1 \cup \mathcal{W}_2 \cup \ldots \cup \mathcal{W}_r$ of $\mathcal{W}$.

STEP 2. We next show that X is locally good, in the sense that for each $x \in X$ there is an open subset $U(x)$ of X such that $x \in U(x)$ and $U(x)$ is good.

Proof Consider a fixed $x \in X$. For each $y \in Y$, $(x, y) \in W(y)$ for some $W(y)$ in $\mathcal{W}$, since $\mathcal{W}$ covers $X \times Y$. By Proposition 10.20 there exist sets $U(y)$, $V(y)$ open in X, Y such that $(x, y) \in U(y) \times V(y) \subseteq W(y)$. The family $\{V(y) : y \in Y\}$ is an open cover for Y, so by compactness of Y there exists a finite subcover, say $\{V(y_1), V(y_2), \ldots, V(y_r)\}$. Let

$$U(x) = U(y_1) \cap U(y_2) \cap \ldots \cap U(y_r).$$

The idea of this move is illustrated schematically in Figure 13.3. For each i in $\{1, 2, \ldots, r\}$ we have

$$U(x) \times V(y_i) \subseteq U(y_i) \times V(y_i) \subseteq W(y_i), \quad \text{so}$$

$$U(x) \times Y = U(x) \times \bigcup_{i=1}^{r} V(y_i) = \bigcup_{i=1}^{r} (U(x) \times V(y_i)) \subseteq \bigcup_{i=1}^{r} W_{y_i}$$

so $U(x)$ is good. Also, $x \in U(x)$ and $U(x)$ is open in X, as a finite intersection of open sets. □

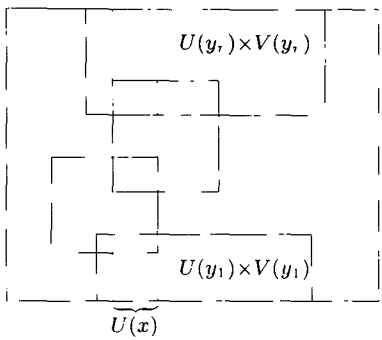

Figure 13.3. Construction of a good neighbourhood of x

STEP 3. Finally we use compactness of X to pass from the local to the global in X. For each $x \in X$ let $U(x)$ be a good open set in X with $x \in U(x)$, as provided by Step 2. Then the family $\{U(x) : x \in X\}$ is an open cover of X, so by compactness of X there is a finite subcover say $\{U(x_1), U(x_2), \ldots, U(x_s)\}$. Since each $U(x_j)$ is good, so is their union by STEP 1, and this is all of X. So X is good, as required. □

Remark Theorem 13.21 extends easily by induction to the product of any finite number of compact spaces. It is true also for infinite products of compact spaces. In its general form it is referred to as Tychonoff's Theorem, and has many applications, for example in analysis. Since we restrict to finite products in this book, we refer to Willard (2004) for the general case.

Compact subsets of Euclidean spaces

Theorem 13.22 (Heine–Borel theorem) *Any closed bounded subset C of $\mathbb{R}^n$ is compact.*

As an illustration, we prove this first for $n = 1$. Suppose that $C \subseteq \mathbb{R}$ is bounded and closed in $\mathbb{R}$. Since C is bounded, $C \subseteq [a, b]$ for some $a, b \in \mathbb{R}$. Now C is closed in $[a, b]$ by Exercise 10.5 and $[a, b]$ is compact by Theorem 13.9, so C is compact as a closed subset of a compact space (Proposition 13.20). We note that compactness of $[a, b]$ is both a special case, and the basis of the proof, of this one-dimensional Heine–Borel theorem.

Proof Suppose that $C \subset \mathbb{R}^n$ is bounded and closed in $\mathbb{R}^n$. Since C is bounded, it is contained in the n-fold product $[a, b]^n$ of some closed

bounded interval $[a, b]$ by Exercise 5.7. Since $[a, b]$ is compact by Theorem 13.9, so is its n-fold product by induction using Theorem 13.21. Also, C is closed in $[a, b]^n$ by Exercise 10.5, so C is compact as a closed subspace of a compact space. □

Remark In more general metric spaces, a subset may be bounded and closed without being compact—as a simple example consider $(0, 1)$ as a subset of itself. There are more interesting examples in the web site, where compact subsets of function spaces are examined.

Compactness and uniform continuity

Definition 13.23 *A map $f : X \to Y$ of metric spaces X, Y with metrics d_X, d_Y is said to be* uniformly continuous *on X if given $\varepsilon > 0$ there exists $\delta > 0$ such that $d_Y(f(x), f(a)) < \varepsilon$ for any $x, a \in X$ satisfying $d_X(x, a) < \delta$.*

Notice that this is stronger than ordinary continuity in that δ can depend on ε but not on a; that is the significance of the word *uniformly*. Ordinary continuity of f is a local property in that it says something about the behaviour of f in some neighbourhood of each point in X. Uniform continuity is a global property since it says something about the behaviour of f over the whole space X. Since compactness allows us to pass from the local to the global, the next proposition is not surprising.

Proposition 13.24 *If $f : X \to Y$ is a continuous map of metric spaces and X is compact then f is uniformly continuous on X.*

For the proof of this result and more on uniform continuity we refer to the web site. However, we note one further result about uniform continuity here.

Proposition 13.25 *If metrics d_1, d_2 for a set X are Lipschitz equivalent, then the identity map of (X, d_1) to (X, d_2) is uniformly continuous as is its inverse.*

The proof is Exercise 13.16.

An inverse function theorem

We end this chapter with a type of inverse function theorem which will be used several times later on.

Proposition 13.26 *Suppose that $f : X \to Y$ is a continuous one–one correspondence, where X is a compact space and Y is a Hausdorff space. Then f is a homeomorphism.*

Proof Since f is injective and onto, we know that there is an inverse function $f^{-1} : Y \to X$; we just have to prove that f^{-1} is continuous. Suppose that V is closed in X. It is enough to show that $(f^{-1})^{-1}(V)$ is closed in Y. By Proposition 3.20 $(f^{-1})^{-1}(V) = f(V)$.

Now the rest of the argument follows from three implications:

$$V \text{ closed in } X \Rightarrow V \text{ is compact } \Rightarrow f(V) \text{ is compact } \Rightarrow f(V) \text{ is closed in } Y.$$

These implications follow because a closed subset of a compact space is compact (Proposition 13.20), the continuous image of a compact space is compact (Proposition 13.15) and a compact subspace of a Hausdorff space is closed (Proposition 13.12). So $(f^{-1})^{-1}(V) = f(V)$ is closed in Y as required. □

Corollary 13.27 *If $f : X \to Y$ is a continuous injective map from a compact space X into a Hausdorff space Y, then f determines a homeomorphism of X onto $f(X)$.*

This follows from Proposition 13.26 and Proposition 10.6.

Example 13.28 If $f : [a, b] \to \mathbb{R}$ is a continuous monotonic function, then it has a continuous inverse function $f^{-1} : f([a, b]) \to [a, b]$ which is also monotonic.

Exercise 13.1 Prove that any indiscrete space is compact.

Exercise 13.2 Prove that a discrete space is compact iff it is finite.

Exercise 13.3 Show that if A and B are compact subsets of a space X then so is $A \cup B$.

Exercise 13.4 Which of the following subsets of $\mathbb{R}$, $\mathbb{R}^2$ are compact?

(i) $[0, 1)$;
(ii) $[0, \infty)$;
(iii) $\mathbb{Q} \cap [0, 1]$;
(iv) $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$;
(v) $\{(x, y) \in \mathbb{R}^2 : |x| + |y| \leqslant 1\}$;

(vi) $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$;
(vii) $\{(x, y) \in \mathbb{R}^2 : x \geqslant 1, 0 \leqslant y \leqslant 1/x\}$.

Exercise 13.5 Given topologies $\mathcal{T}, \mathcal{T}'$ on a set X with $\mathcal{T} \subseteq \mathcal{T}'$, prove that if $(X, \mathcal{T}')$ is compact then so is $(X, \mathcal{T})$.

Exercise 13.6 Prove that the following is a necessary and sufficient condition for a space X to be compact: if $\{V_i : i \in I\}$ is any indexed family of closed subsets of X such that $\bigcap_{j \in J} V_j$ is non-empty for any finite subset $J \subseteq I$ then $\bigcap_{i \in I} V_i$ is non-empty.
[Hint: Take complements and apply the definition of compactness.]

Exercise 13.7 Obtain another proof that finite subsets of a Hausdorff space are closed using the result that a compact subset of a Hausdorff space is closed.

Exercise 13.8 Prove that if $X \subseteq \mathbb{R}$ is not compact, then there is a continuous function $f : X \to \mathbb{R}$ which is not bounded.
[Hint: Consider separately the cases X is not bounded and X is not closed in $\mathbb{R}$.]

Exercise 13.9 Prove that if $X \subseteq \mathbb{R}$ is not compact, then there is a continuous function $f : X \to \mathbb{R}$ which is bounded but does not attain its bounds.

Exercise 13.10 Prove that if C, C' are compact subsets of a Hausdorff space X then $C \cap C'$ is compact.

Exercise 13.11 Suppose for every $n \in \mathbb{N}$ that V_n is a non-empty closed subset of a compact space X with $V_n \supseteq V_{n+1}$. Prove that $V_\infty = \bigcap_{n=1}^{\infty} V_n \neq \emptyset$.

Exercise 13.12* With the notation of Exercise 13.11, suppose that X is as compact. Suppose that U is open in X and that $V_\infty \subseteq U$. Prove that $V_n \subseteq U$ for some $n \in \mathbb{N}$.
[Hint: Consider the sequence of sets $W_n = V_n \cap (X \setminus U)$.]

Exercise 13.13* Suppose that X is a compact Hausdorff space and suppose that $f : X \to X$ is a continuous map. Let $X_0 = X$, $X_1 = f(X_0)$ and inductively define $X_{n+1} = f(X_n)$ for $n \geqslant 1$.

(a) Show that $A = \bigcap_{n=0}^{\infty} X_n$ is non-empty. [Hint. Remember Exercise 13.11.]

(b) Show further that $f(A) = A$. [Hint: To show that $a \in A$ is in $f(A)$ apply Exercise 13.11 to the sets $V_n = f^{-1}(a) \cap X_n$.]

Exercise 13.14 Suppose that X is a compact metric space with metric d and that $f : X \to X$ is a continuous map such that for every $x \in X$, $f(x) \neq x$ Prove that there exists $\varepsilon > 0$ such that $d(f(x), x) \geqslant \varepsilon$ for all $x \in X$.
[Hint: Show that the map $g : X \to \mathbb{R}$ defined by $g(x) = d(f(x), x)$ is continuous so attains its bounds.]

Exercise 13.15 (Bisection method) Give another proof of the one-dimensional Heine-Borel theorem along the following lines Let $\mathcal{U}$ be any open cover of $[a, b]$. Assume for a contradiction that there is no finite subcover, and put $c_1 = (a+b)/2$. Then at least one of $[a, c_1]$, $[c_1, b]$ is not covered by any finite subfamily of $\mathcal{U}$. Let $[a_1, b_1]$ denote one of those subintervals which is not so covered. Repeat this argument inductively to get two real sequences (a_n), (b_n) such that for all $n \in \mathbb{N}$.

(i) $a_n \leqslant a_{n+1} < b_{n+1} \leqslant b_n$,
(ii) $b_n - a_n = (b - a)/2^n$;
(iii) $[a_n, b_n]$ is not covered by any finite subfamily of $\mathcal{U}$

Deduce that (a_n), (b_n) both converge to some $c \in [a, b]$ and get a contradiction by showing that for sufficiently large n the interval $[a_n, b_n]$ is contained in a single set U from $\mathcal{U}$.

Exercise 13.16 Suppose that metrics d_1, d_2 for a set X are Lipschitz equivalent (Definition 6 33). Prove that the identity map of X is uniformly (d_1, d_2)-continuous and also uniformly (d_2, d_1)-continuous.

Exercise 13.17 Use the inverse function theorem above to prove that if $\mathcal{T}_1$, $\mathcal{T}_2$ are topologies on a set X such that $\mathcal{T}_1 \subseteq \mathcal{T}_2$ and the spaces $(X, \mathcal{T}_1)$, $(X, \mathcal{T}_2)$ are respectively Hausdorff and compact, then $\mathcal{T}_1 = \mathcal{T}_2$. Deduce that there is no Hausdorff topology on $[0, 1]$ which is strictly coarser than the Euclidean one.

Exercise 13.18 ★ Suppose that $\mathcal{F}$ is a family of real-valued continuous functions defined on a compact space X, with the properties.

(i) $f(x) \geqslant 0$ for all $x \in X$ and all $f \in \mathcal{F}$,
(ii) if $f, g \in \mathcal{F}$ then $f + g \in \mathcal{F}$:
(iii) for each $f \in \mathcal{F}$ there is some point $x_f \in X$ such that $f(x_f) = 0$

Prove that there is some point $x_0 \in X$ such that $f(x_0) = 0$ for all $f \in \mathcal{F}$.
[Hint: argue by contradiction.]

Exercise 13.19 ★ Prove that any compact Hausdorff space is regular. Then prove that it is normal (See Definition 11.8.)

Exercise 13.20 ★

(a) Suppose that $p_X : X \times Y \to X$ is the projection map on the first factor of a topological product $X \times Y$, where Y is compact. Prove that if $W \subseteq X \times Y$ is closed in $X \times Y$ then $p_X(W)$ is closed in X.

(b) Give an example to show that in general (a) is false if we omit the hypothesis that Y is compact

Exercise 13.21 Suppose that $f : X \to Y$ is a map from a space X to a compact space Y and that its graph G_f is closed in $X \times Y$. Prove that f is continuous. [Hint: for any subset $V \subseteq Y$ prove that $f^{-1}(V) = p_X(G_f \cap p_Y^{-1}(V))$. Apply this to V closed in Y, using Proposition 9.5. Use also Exercise 13.20(a).]

Exercise 13.22 ★ For any topological space $(X, \mathcal{T})$. let $X' = X \cup \{\infty\}$ where ∞ is any object not in X. Let $\mathcal{T}'$ be the union of $\mathcal{T}$ with all sets of the form $V \cup \{\infty\}$ where $V \subseteq X$ and $X \setminus V$ is compact and closed in $(X, \mathcal{T})$. Prove that $(X', \mathcal{T}')$ is a compact space containing $(X, \mathcal{T})$ as a subspace.
(Then $(X', \mathcal{T}')$ is called the Alexandroff one-point compactification of $(X, \mathcal{T})$.)

14 Sequential compactness

We recall the form of Bolzano Weierstrass theorem stated in Chapter 4: any bounded sequence of real numbers has at least one convergent subsequence. (You may have seen an equivalent form, which says that any bounded infinite subset of real numbers has at least one limit point; the sequence version is more relevant to our purposes.) The Bolzano–Weierstrass theorem is closely related to the one-dimensional Heine–Borel theorem. For example, we have proved that a continuous real-valued function on a closed bounded interval $[a, b]$ is bounded and attains its bounds using the Heine–Borel theorem (more precisely, the special case for $[a, b]$); equally well this can be proved using the Bolzano–Weierstrass theorem—in analysis textbooks it is often proved that way, for example see Theorems 4.3.1 and 4.3.2 of Hart (2001).

In this chapter we generalize the Bolzano–Weierstrass property of real numbers to give an alternative approach to compactness called sequential compactness. Sequential compactness is often useful in analysis, both for theoretical purposes, proving existence theorems for solutions of certain problems, and also for practical purposes, giving numerical approximations to solutions.

We discuss sequential compactness exclusively for metric spaces, although the concept makes sense more generally. Later in the chapter we pin down and generalize the connection between the Heine–Borel theorem and the Bolzano–Weierstrass theorem by showing that for any metric space sequential compactness is equivalent to compactness in the sense of the previous chapter.

Sequential compactness for real numbers

We begin with a few examples related to the Bolzano–Weierstrass theorem.

Example 14.1 Consider the following sequences of real numbers:

(a) $1, 0, 1, 0, 1, \ldots$;

(b) $1, 0, 2, 0, 3, 0, \ldots$;

(c) $1, 2, 3, 4, \ldots$

The sequence in (a) is bounded but not convergent. However, there are many convergent subsequences: any subsequence which eventually settles down to taking either odd terms only, or even terms only, is convergent. We note that the existence of at least one convergent subsequence is guaranteed by the Bolzano–Weierstrass theorem.

The sequence in (b) is unbounded, but we can still find convergent subsequences—any subsequence which eventually settles down to pick out only zero terms is convergent. So an unbounded sequence can easily have convergent subsequences.

However, in (c) no subsequence is bounded so no subsequence is convergent.

Definition 14.2 *A subset $S \subseteq \mathbb{R}$ is called* sequentially compact *if every sequence in S has at least one subsequence converging to a point in S.*

The Bolzano Weierstrass theorem in Chapter 4 says that any bounded sequence of real numbers has a convergent subsequence. From this we easily deduce

Proposition 14.3 *Any closed bounded subset $S \subseteq \mathbb{R}$ is sequentially compact.*

Proof Let (x_n) be any sequence in S. By Theorem 4.19 there is at least one convergent subsequence (x_{n_r}). Since S is closed in $\mathbb{R}$ the limit of (x_{n_r}) is in S (by Corollary 6.30). So S is sequentially compact. □

The converse is also true.

Proposition 14.4 *Any sequentially compact subset of $\mathbb{R}$ is closed and bounded.*

This is a special case of Exercises 14.2 and 14.4, and we omit the proof. The next theorem is an immediate consequence of Proposition 14.3 and Proposition 14.4.

Theorem 14.5 *A subset $S \subseteq \mathbb{R}$ is sequentially compact iff it is bounded and closed in $\mathbb{R}$.*

We showed in Chapter 13 show that a subset of $\mathbb{R}$ is compact iff it is bounded and closed in $\mathbb{R}$ (this is the one-dimensional Heine–Borel theorem); with Theorem 14.5 this gives the following result.

Theorem 14.6 *A subset of $\mathbb{R}$ is compact iff it is sequentially compact.*

Sequential compactness for metric spaces

The next definition generalizes the notion of sequential compactness to metric spaces.

Definition 14.7 *A metric space X is* sequentially compact *if every sequence in X has at least one subsequence converging to a point of X.*

Definition 14.8 *A non-empty subset A of a metric space (X, d) is* sequentially compact *if, with the subspace metric d_A, it satisfies Definition 14.7. Conventionally the empty set is considered to be sequentially compact.*

Example 14.9 (a) Any finite metric space is sequentially compact.
(b) Any bounded closed subset of $\mathbb{R}$ is sequentially compact.

Proof (a) Given any sequence in a finite metric space X, at least one point of X must be repeated infinitely often in the sequence. The occurrences of such an infinitely repeated point give a constant, hence a convergent, subsequence.
(b) This is Proposition 14.3. □

In the rest of this chapter we prove the following generalization of Theorem 14.6. Its proof is slightly sophisticated—for example, some of it proceeds by contradiction rather than direct construction. We shall illustrate some of the moves with related examples.

Theorem 14.10 *A metric space is compact iff it is sequentially compact.*

Remark For subsets of $\mathbb{R}$ this is Theorem 14.6.
As we have mentioned, sequential compactness makes sense more generally, and in fact the analogue of Theorem 14.10 holds in some spaces which are not metrizable, but this is beyond our scope; see for example 17G in Willard (2004).

Towards proving Theorem 14.10 we first show that a compact metric space is sequentially compact. For this we use the following result.

Proposition 14.11 *Let (x_n) be a sequence in a metric space X and let $x \in X$. Suppose that for each $\varepsilon > 0$ the neighbourhood $B_\varepsilon(x)$ contains x_n for infinitely many values of n. Then (x_n) has a subsequence converging to x.*

Example 14.12 Consider the following real sequence

$$1, 1, 2, 1/2, 3, 1/3, \ldots, n, 1/n, \ldots$$

Then 0 is such that for any $\varepsilon > 0$ the neighbourhood $B_\varepsilon(0)$ contains x_n for infinitely many n, since it contains every term of the form $1/n$ in

the sequence for n sufficiently large (namely, $n > 1/\varepsilon$). The subsequence of (x_n) formed by taking every second term is $(1/n)$ which converges to 0.

Proof of 14.11. Let n_1 be such that $x_{n_1} \in B_1(x)$. Suppose inductively that positive integers $n_1 < n_2 < \ldots < n_r$ have been chosen so that $x_{n_i} \in B_{1/i}(x)$ for each $i = 1, 2, \ldots, r$. Since $B_{1/(r+1)}(x)$ contains x_n for infinitely many values of n, it must contain x_n for some $n > n_r$; we take n_{r+1} to be such an n. This inductive procedure shows that (x_n) has a subsequence (x_{n_r}) with $x_{n_r} \in B_{1/r}(x)$ for all r, so this subsequence converges to x. □

NB Notice that the condition says that $B_\varepsilon(x)$ contains x_n 'for infinitely many values of n', not that it contains 'infinitely many different points in the set $\{x_n : n \in \mathbb{N}\}$'. In particular, (x_n) might be the sequence $(x, x, x, \ldots, x, \ldots)$ and this would satisfy the hypotheses of the proposition.

Corollary 14.13 *Suppose that a sequence (x_n) in a metric space X has no convergent subsequences. Then for each $x \in X$ there exists $\varepsilon_x > 0$ such that $B_{\varepsilon_x}(x)$ contains x_n for only finitely many values of n.*

Example 14.14 (a) In $\mathbb{R}$ consider the sequence (n), which has no convergent subsequence. Now for any $x \in \mathbb{R}$ we may take $\varepsilon_x = 1$ and the neighbourhood $B_{\varepsilon_x}(x)$ contains at most two terms in the sequence (n) (it contains only one term when x is an integer).

(b) In the metric space $X = (0, 1]$ consider the sequence $(1/n)$. Every subsequence is 'trying to converge to 0', which is not in X. So $(1/n)$ has no convergent subsequences in X. For any $x \in X$ we have $0 < x \leqslant 1$, so $1/(N+1) \leqslant x \leqslant 1/N$ for some integer $N > 0$. Let us now choose $\varepsilon_x = 1/(N+1) - 1/(N+2)$. Then $\varepsilon_x > 0$ and $x - \varepsilon_x \geqslant 1/(N+2)$. Hence $B_{\varepsilon_x}(x)$ contains $1/n$ only for $n \leqslant N+1$, so only finitely many terms of the sequence $(1/n)$. (In fact with some more effort one can show that it contains at most two terms of the sequence.)

Now we can prove

Theorem 14.15 *Any compact subset X of a metric space Y is sequentially compact.*

Proof Suppose that (x_n) is a sequence in compact X and suppose for a contradiction that (x_n) has no convergent subsequence. By Corollary 14.13, for every $x \in X$ there exists $\varepsilon_x > 0$ such that $B_{\varepsilon_x}(x)$ contains x_n for only finitely many n. The family $\{B_{\varepsilon_x}(x) : x \in X\}$ is an open cover for X, so

there is a finite subcover. Each set in this finite subcover contains x_n for only finitely many values of n. But this implies that the whole of X contains x_n for only finitely many values of n, which is nonsense since (x_n) is a sequence in X. So (x_n) must have a subsequence converging to some point x. Now X is compact and hence closed in Y by Proposition 13.12. So x is in X by Corollary 6.30. Thus X is sequentially compact. □

The proof that a sequentially compact metric space is compact is longer, and it is convenient first to prove two preliminary results.

Definition 14.16 *Let $\mathcal{U}$ be any family of subsets of a metric space X covering a subset $A \subseteq X$. A* Lebesgue number for $\mathcal{U}$ *is a real number ε with $\varepsilon > 0$ such that for any $a \in A$ the ball $B_\varepsilon(a)$ is contained in some single set from $\mathcal{U}$.*

Example 14.17 Consider the open cover of $[0, 1]$ by the sets $(-1, 3/4)$ and $(1/4, 2)$. Then $\varepsilon = 1/4$ is a Lebesgue number for this cover. For let $x \in [0, 1]$. If $0 \leqslant x \leqslant 1/2$ then $x + 1/4 \leqslant 3/4$ and it follows that $(x - 1/4, x + 1/4) \subseteq (-1, 3/4)$, while if $1/2 \leqslant x \leqslant 1$ then $x - 1/4 \geqslant 1/4$ and $(x - 1/4, x + 1/4) \subseteq (1/4, 2)$.

Proposition 14.18 *Any open cover $\mathcal{U}$ of a sequentially compact metric space X has a Lebesgue number.*

If ε is a Lebesgue number for $\mathcal{U}$ then so is any δ with $0 < \delta \leqslant \varepsilon$. The proof of Proposition 14.18 is by contradiction; it gives no idea how to find a specific Lebesgue number. (A more direct proof, but one which works only for finite covers, is developed in Exercise 14.14.)

Proof Suppose for a contradiction that there is no such ε. Then in particular, for any positive integer n the number $1/n$ is not a Lebesgue number for $\mathcal{U}$. So there exists some point in X, which we shall call x_n, such that $B_{1/n}(x_n)$ is not contained in any single set of $\mathcal{U}$. By sequential compactness of X, the sequence (x_n) has some subsequence (x_{n_r}) converging to a point $x \in X$. Now $x \in U$ for some $U \in \mathcal{U}$, and since U is open in X, there exists $\varepsilon > 0$ such that $B_{2\varepsilon}(x) \subseteq U$. By convergence of (x_{n_r}) to x, there exists $R \in \mathbb{N}$ such that $x_{n_r} \in B_\varepsilon(x)$ for all $r \geqslant R$. In particular, we may choose $r \geqslant R$ large enough so that $1/n_r < \varepsilon$. It is now sufficient to prove that $B_{1/n_r}(x_{n_r}) \subseteq B_{2\varepsilon}(x)$, for then $B_{1/n_r}(x_{n_r}) \subseteq U$, contradicting the choice of x_{n_r}.

Suppose that $y \in B_{1/n_r}(x_{n_r})$. Then

$d(y, x) \leqslant d(y, x_{n_r}) + d(x_{n_r}, x) < 1/n_r + \varepsilon < 2\varepsilon$, so $y \in B_{2\varepsilon}(x)$ as required.

□

The other result we need also requires a definition.

Definition 14.19 *Given a real number $\varepsilon > 0$ and a metric space X, a subset $N \subseteq X$ is called* an ε-net for X *if the family $\{B_\varepsilon(x) : x \in N\}$ covers X.*

Example 14.20 The set of integer lattice points in the plane (the points both of whose coordinates are integers) is a 1-net for the plane.

This indicates the reason for the name *net*. In fact the integer lattice points form an ε-net for the plane for any $\varepsilon > 1/\sqrt{2}$, since any point in the plane is at distance at most $1/\sqrt{2}$ from the nearest lattice point. However, the interesting ε-nets for us are the finite ones.

Proposition 14.21 *Let (X, d) be a sequentially compact metric space, and let $\varepsilon > 0$. Then there exists a finite ε-net for X.*

Proof Suppose for a contradiction that there is some $\varepsilon > 0$ for which X has no finite ε-net. We shall construct a sequence in X with no convergent subsequence, contradicting sequential compactness of X. Let x_1 be any point of X. Then $\{x_1\}$ is not an ε-net for X, so there exists $x_2 \in X$ with $d(x_2, x_1) \geqslant \varepsilon$. Suppose inductively that $x_1, x_2, \ldots, x_n$ have been chosen in X such that $d(x_i, x_j) \geqslant \varepsilon$ whenever $i, j \in \{1, 2, \ldots, n\}$ and $i \neq j$. Since $\{x_1, x_2, \ldots, x_n\}$ is not an ε-net for X there exists a point in X, call it x_{n+1}, such that $d(x_{n+1}, x_j) \geqslant \varepsilon$ for $j = 1, 2, \ldots, n$. This inductive procedure gives us a sequence (x_n) in X such that $d(x_m x_n) \geqslant \varepsilon$ whenever $m \neq n$. So (x_n) has no Cauchy subsequence and hence by Proposition 6.28 no convergent subsequence. □

Theorem 14.22 *Any sequentially compact metric space X is compact.*

Proof Suppose that $\mathcal{U}$ is an open cover of a sequentially compact metric space X. By Proposition 14.18 there is a Lebesgue number $\varepsilon > 0$ for $\mathcal{U}$. By Proposition 14.21 there is a finite ε-net, say $\{x_1, x_2, \ldots, x_r\}$ for X. By definition of Lebesgue number, for each $i = 1, 2, \ldots, r$ there is some single set, call it U_i of $\mathcal{U}$ such that $B_\varepsilon(x_i) \subseteq U_i$. Then

$$X \subseteq \bigcup_{i=1}^{r} B_\varepsilon(x_i) \subseteq \bigcup_{i=1}^{r} U_i,$$

so $\mathcal{U}$ has the finite subcover $\{U_1, U_2, \ldots, U_r\}$ for X, showing X to be compact. □

We end this chapter with an example of a function space which is not sequentially compact (involving 'the moving bump' below).

Figure 14.1. Graphs of f_1 and f_2

Example 14.23 Let (f_n) be the sequence of real-valued continuous functions on $[0, 1]$ defined as follows (see Figure 14.1): for any $n \geqslant 1$,

$$f_n(x) = \begin{cases} 0 & \text{if } 0 \leqslant x \leqslant 1/2^n \\ 2^{n+1}(x - 1/2^n) & \text{if } 1/2^n \leqslant x \leqslant 1/2^n + 1/2^{n+1} \\ 2^{n+1}(1/2^{n-1} - x) & \text{if } 1/2^n + 1/2^{n+1} \leqslant x \leqslant 1/2^{n-1} \\ 0 & \text{if } 1/2^{n-1} \leqslant x \leqslant 1 \end{cases}$$

We consider $B = \{f_n : n \in \mathbb{N}\}$ as a subset of $\mathcal{C}[0, 1]$ with the sup metric d_∞. Then we can see that $d_\infty(f_m, f_n) = 1$ whenever $m \neq n$. Hence the sequence (f_n) in B contains no Cauchy subsequence and hence no convergent subsequence. Hence B is not compact, and in fact (see Exercise 14.12 below) it is not relatively compact in $\mathcal{C}[0, 1]$.

The first eleven exercises below could be done by using the equivalence of sequential compactness with compactness and referring to previous results, but it is suggested that they are attempted without using Theorem 14.10, in order to gain familiarity with sequential compactness. The later exercises are designed with the idea that sequential compactness may be the more appropriate way to tackle them.

Exercise 14.1 Prove that the open unit interval $(0, 1)$ is not sequentially compact.

Exercise 14.2 Prove that a sequentially compact metric space is bounded

Exercise 14.3 Prove that a closed subset of a sequentially compact metric space is sequentially compact.

Exercise 14.4 Prove that a sequentially compact subspace of a metric space X is closed in X.

Exercise 14.5 Prove that if $f : X \to Y$ is a continuous map of metric spaces and X is sequentially compact then so is $f(X)$.

Exercise 14.6 Prove that if metric spaces X_1 and X_2 are homeomorphic then X_1 is sequentially compact iff X_2 is sequentially compact.

Exercise 14.7 Prove that any continuous map from a sequentially compact metric space to another metric space has bounded image.

Exercise 14.8 Suppose that X is a sequentially compact metric space and that $f : X \to \mathbb{R}$ is continuous. Prove that f attains its bounds on X.

Exercise 14.9 Prove that the product of two sequentially compact metric spaces is sequentially compact.

Exercise 14.10 Prove that a closed bounded subset of $\mathbb{R}^n$ is sequentially compact. (You may assume the result for $n = 1$.)

Exercise 14.11 Suppose that X is a sequentially compact metric space and we are given a nested sequence $V_1 \supseteq V_2 \supseteq \ldots$ of non-empty closed subsets of X. Prove that $\bigcap_{n=1}^{\infty} V_n \neq \emptyset$.

Exercise 14.12 Prove that a subspace C of a metric space X is relatively compact in X iff every sequence in C has a convergent subsequence.

Exercise 14.13 Give another proof that $[a, b]$ is connected along the following lines. Suppose that $\{A, B\}$ is a partition of $[a, b]$, where $a \in A$. Since A, B are open in $[a, b]$, for each $x \in [a, b]$ there is some $B_{\delta(x)}(x)$ entirely contained in either A or B. Let ε be a Lebesgue number for the open cover $\{B_{\delta(x)}(x) : x \in [a, b]\}$ of $[a, b]$ and let $a = a_0 < a_1 < \ldots < a_n = b$ be such that $a_i - a_{i-1} < \varepsilon$ for $i = 1, 2, \ldots, n$. Deduce that $[a, b] \subseteq A$.

Exercise 14.14* Suppose that X is a sequentially compact metric space and that $\mathcal{U} = \{U_1, U_2, \ldots, U_n\}$ is an open cover of X. Put $C_i = X \setminus U_i$. Show that if $U_i = X$ for some $i \in \{1, 2, \ldots, n\}$ then any $\varepsilon > 0$ is a Lebesgue number for $\mathcal{U}$. From now on, assume that no U_i is X.

(i) Recalling that $d(x, C_i) = \inf\{d(x, c) : c \in C_i\}$, prove that $x \mapsto d(x, C_i)$ is a continuous real-valued function on X all of whose values are non-negative.

(ii) Defining $f(x) = \dfrac{1}{n}\displaystyle\sum_{i-1}^{n} d(x, C_i)$, show that $f : X \to \mathbb{R}$ is continuous with a positive value for each $x \in X$.

(iii) Using Exercise 14.8, deduce that there exists $\varepsilon > 0$ such that $f(x) \geqslant \varepsilon$ for all $x \in X$.

(iv) For any $x \in X$ show that $f(x) \leqslant \max\{d(x, C_i) : i \in \{1, 2, \ldots, n\}\}$.

(v) For any $x \in X$, show that $B_\varepsilon(x) \subseteq U_{k(x)}$ where $k(x) \in \{1, 2, \ldots, n\}$ is such that $d(x, C_{k(x)}) = \max\{d(x, C_i) : i \in \{1, 2, \ldots, n\}\}$. Deduce that ε is a Lebesgue number for $\mathcal{U}$.

Exercise 14.15* Suppose that X is a compact metric space and $V_1 \supseteq V_2 \supseteq \ldots$ is a nested sequence of closed subsets of X. Prove that

$$\operatorname{diam}\left(\bigcap_{n=1}^{\infty} V_n\right) = \inf\{\operatorname{diam} V_n : n \in \mathbb{N}\}.$$

Exercise 14.16 Let f_n be as in Example 14.23 above, and for each $n \in \mathbb{N}$ let V_n be the set $\{f_m : m \geqslant n\}$.

(a) Prove that $\bigcap_{n=1}^{\infty} V_n = \emptyset$.

(b) Prove that $\operatorname{diam} V_n = 1$ for any n.

(c) Deduce that the conclusion of Exercise 14.15 fails in this case.

Exercise 14.17** (a) Let X be a compact metric space with metric d and suppose that $f : X \to X$ satisfies $d(f(x), f(y)) = d(x, y)$ for all $x, y \in X$. Prove that f is onto (so f is an isometry).

(b) Let X, Y be compact metric spaces with metrics d_X, d_Y. Suppose that the maps $f : X \to Y$ and $g : Y \to X$ satisfy $d_Y(f(x_1), f(x_2)) = d_X(x_1, x_2)$ and $d_X(g(y_1), g(y_2)) = d_Y(y_1, y_2)$ for all $x_1, x_2 \in X$ and all $y_1, y_2 \in Y$. Prove that f and g are both onto (and hence isometries).

(c) Construct a function $f : (0, \infty) \to (0, \infty)$ which is not onto but nevertheless satisfies $|f(x) - f(y)| = |x - y|$ for all $x, y \in (0, \infty)$.

15 Quotient spaces and surfaces

It would have been logical to discuss quotient spaces immediately after subspaces and products, but it is convenient before tackling them to have available some of our compactness results. Also, quotients are sometimes found more challenging than subspaces and products, although they are close to popular expositions of topology involving Möbius bands, doughnuts, and such geometric objects. This chapter will give a basic account of quotient spaces, and at the same time consider a few standard surfaces. There are further results about quotients on the web site.

Motivation

At this point it would be good to get a sheet of newspaper, a pair of scissors and some paste, and construct a Möbius band: Take a rectangular strip of paper, twist one of the shorter ends through 180° and paste it to the other short end (see Figure 15.1(a)). As the reader probably knows, the Möbius band has some unusual properties. For example, try cutting it parallel to a longer edge of the original rectangular strip and halfway across the band, and keep cutting until you get back to where you started: In the end instead of getting two pieces as you might expect, you get a single twisted cylindrical strip, twice the length of the Möbius band. On the other hand, if you cut parallel to a longer edge of the original rectangle and one-third of the way across the band, and keep cutting until you get back to where you started, the band falls apart into two interlinked pieces; the shorter is a Möbius band and the longer is a twisted cylinder. (Although we shall not prove it, a band with an even number of twists is homeomorphic to a band with no twists, hence the name 'twisted cylinder'.)

We can represent in a diagram how we have constructed a Möbius band: in Figure 15.1(b) the shorter edges are stuck together in the way that the arrows indicate.

Next take a rectangle of paper again, this time in your mind or in a diagram. First stick the longer edges together as indicated by the arrows in Figure 15.2 to get a cylinder. Now stick together the shorter edges as the arrows suggest. What we get is like the surface of a doughnut, which

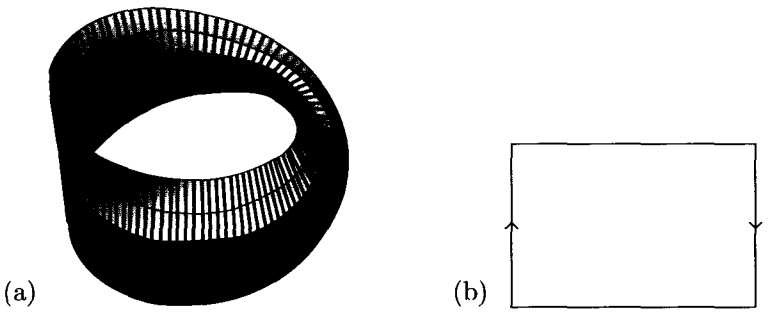

Figure 15.1. (a) Möbius band and (b) schematic Möbius band

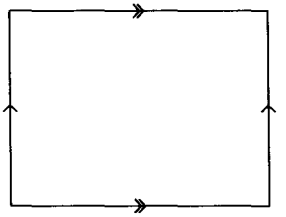

Figure 15.2. Torus

is called a *torus* (the name comes from Greek architecture). This is one of the standard surfaces which we study in more detail later.

Next, suppose we have a pentagon as in Figure 15.3(a) and we stick the edges together in the way the arrows in Figure 15.3(b) indicate. Notice that after we have done the sticking, the vertex a gets stuck to e which gets stuck to c. Also, a gets stuck to b which gets stuck to d. We often label these vertices as they appear after sticking, so all the vertices in Figure 15.3(b) get labelled a. This is like a torus except that there is a 'free' edge going from the vertex a back to itself. So another way of drawing it is as in Figure 15.4(a). We see that what we have is a torus with a hole in it, see Figure 15.4(b).

Figure 15.3. (a) Pentagon and (b) pentagon with edges stuck together

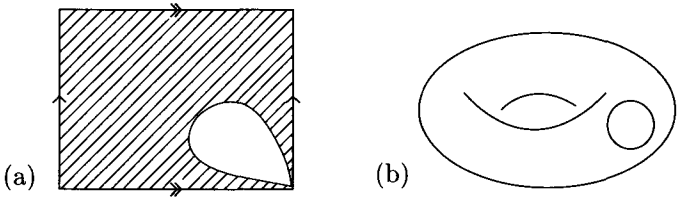

Figure 15.4. (a) Another representation and (b) torus with one hole

In all of these examples we have started with one topological space, namely a subspace of $\mathbb{R}^2$, and stuck some subsets of it together to get another topological space. This is essentially the idea of quotient spaces.

A formal approach

It is a matter of discussion as to whether diagrams such as Figure 15.2 constitute 'rigorous' mathematics. Many people find such diagrams geometrically illuminating and consider it easier to follow what is going on from them than from lists of formulae—for further examples see Figure 15.10 below. However, in order to put our work in line with the previous treatments of subspaces, products, and so on, we now consider a more formal approach. It is time to introduce some terminology. The mathematical term for 'sticking things together' is 'identifying'. For example, to get a cylinder we begin with a square as in Figure 15.5 and 'identify ab with dc', meaning that each pair such as $\{x, x'\}$ in Figure 15.5 is to be thought of as just *one* point in the space that we are constructing. The way to formalize this idea of identifying edges is via *equivalence classes* (as for quotient structures in group theory). Given any topological space X and

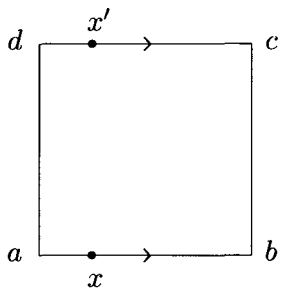

Figure 15.5. Cylinder

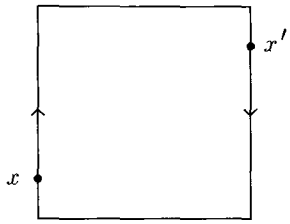

Figure 15.6. Points to be identified for the Möbius band

an equivalence relation $\sim$ on its points, we form a new space $X/\sim$ whose points are the equivalence classes in X. (Thus there is just one point in $X/\sim$ corresponding to all the points in a given equivalence class in X.) This is well illustrated by the following example.

Example 15.1 We consider the Möbius band again. Formally we can define an equivalence relation on $X = [0, 1] \times [0, 1] \subseteq \mathbb{R}^2$, as follows.

For (s_1, t_1), (s_2, t_2) in X, let $(s_1, t_1) \sim (s_2, t_2)$ iff one of the following holds:

(i) $s_1 = s_2$ and $t_1 = t_2$;
(ii) $s_1 = 0$, $s_2 = 1$, and $t_2 = 1 - t_1$;
(iii) $s_1 = 1$, $s_2 = 0$, and $t_2 = 1 - t_1$.

We can bracket (ii) and (iii) together into one condition:

(iv) $\{s_1, s_2\} = \{0, 1\}$ and $t_2 = 1 - t_1$.

It is straightforward to check that this is an equivalence relation. For example, to prove symmetry, suppose that $(s_1, t_1) \sim (s_2, t_2)$. Then one of (i), (iv) above holds. If $s_1 = s_2$ and $t_1 = t_2$, then $s_2 = s_1$ and $t_2 = t_1$ so $(s_2, t_2) \sim (s_1, t_1)$. On the other hand if $\{s_1, s_2\} = \{0, 1\}$ and $t_2 = 1 - t_1$, then $\{s_2, s_1\} = \{0, 1\}$ and $t_1 = 1 - t_2$, so $(s_2, t_2) \sim (s_1, t_1)$ by (iv). The proofs of reflexivity and transitivity are Exercise 15.2.

Next is a description of the equivalence classes corresponding to the above equivalence relation. If $0 < s < 1$ then the singleton set $\{(s, t)\}$ is an equivalence class on its own for any fixed $t \in [0, 1]$. But for each $t \in [0, 1]$ the pair $\{(0, t), (1, 1 - t)\}$ is an equivalence class, as it should be since we want to stick these two points together (see Figure 15.6). The proof of this is Exercise 15.3. Now we can see geometrically that $X/\sim$ is at least the same *set* as the Möbius band. When we have examined the appropriate topology for $X/\sim$ later, we shall see that $X/\sim$ is indeed the Möbius band.

Example 15.2 We now consider the torus more formally. We again take $X = [0, 1] \times [0, 1]$, but now let $(s_1, t_1) \sim (s_2, t_2)$ iff one of the following holds:

(i) $s_1 = s_2$ and $t_1 = t_2$;
(ii) $\{s_1, s_2\} = \{0, 1\}$, $t_1 = t_2$;
(iii) $\{t_1, t_2\} = \{0, 1\}$, $s_1 = s_2$;
(iv) $\{s_1, s_2\} = \{0, 1\}$, $\{t_1, t_2\} = \{0, 1\}$.

Again it is straightforward to check that this is an equivalence relation, although it is slightly lengthy due to the number of cases involved. The details are on the web site.

The corresponding equivalence classes are:

$\{(s, t)\}$ for any s, t with $0 < s < 1$ and $0 < t < 1$,
$\{(0, t), (1, t)\}$ for any t with $0 < t < 1$,
$\{(s, 0), (s, 1)\}$ for any s with $0 < s < 1$,
$\{(0, 1), (1, 0), (0, 0), (1, 1)\}$.

Again we can see geometrically that $X/\sim$ is at least the same *set* as the torus, and later when we have topologised $X/\sim$ we shall see that it is indeed the torus.

We end this section with a purely set-theoretic result about maps of quotients.

Proposition 15.3 *Suppose that X, Y are sets and $\sim$ is an equivalence relation on X. Let $f : X \to Y$ be a map such that $f(x) = f(y)$ whenever $x \sim y$. Then there is a well-defined map $g : X/\sim \to Y$ where if $\{x\}$ denotes the equivalence class of x under $\sim$, we define $g(\{x\}) := f(x)$. We say that 'f respects the identifications on X' and we call g the map 'induced by f'.*

Proof To see that g is well defined we need to check that $f(x') = f(x)$ whenever $x' \in \{x\}$. But if $x' \in \{x\}$ then $x' \sim x$ so $f(x') = f(x)$ by assumption. □

The quotient topology

Now given a topological space X and an equivalence relation $\sim$ on it, we want to put a sensible topology on $X/\sim$ related to the topology on X and giving us the topology we know we want on examples like the Möbius band and the torus, coming from their representations as subspaces of

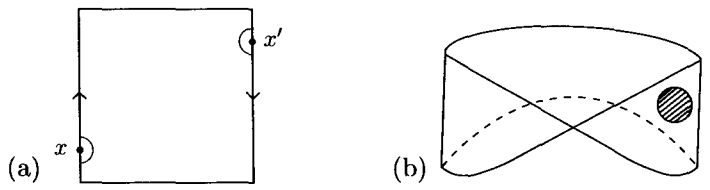

Figure 15.7. (a) Topologizing a Möbius band and (b) open disc in Möbius band

Euclidean three-space. We can get a hint of how to do this from the Möbius band: a 'small round open set' in the Möbius band centred on the point represented by x, x' in Figure 15.7(a) should be the image in the band of something like the union of an open half-disc around x and another around x' as shown. For, if mentally we cut these out and flip one of them over they fit together to give you an open disc (see Figure 15.7(b)) in the Möbius band.

This kind of consideration suggests taking as open sets in $X/\sim$ those whose pre-images 'up in X' are open in X. Next we formalize the preceding.

Proposition 15.4 *Suppose that $(X, \mathcal{T})$ is a topological space and that $\sim$ is an equivalence relation on X. Denote the set of equivalence classes by $X/\sim$, and let $p : X \to X/\sim$ be the function which assigns to each point of X the equivalence class it is in. Let $\tilde{\mathcal{T}}$ be the family of all subsets $\tilde{U} \subseteq X/\sim$ such that $p^{-1}(\tilde{U}) \in \mathcal{T}$. Then $\tilde{\mathcal{T}}$ is a topology for $X/\sim$, called the* quotient *topology, $(X/\sim, \tilde{\mathcal{T}})$ is called a* quotient space *(of X), and $p : X \to X/\sim$ is called the* natural map.

Proof (T1) $p^{-1}(X/\sim) = X \in \mathcal{T}$ so $X/\sim \in \tilde{\mathcal{T}}$. Also, $p^{-1}(\emptyset) = \emptyset$ which is in $\mathcal{T}$ so $\emptyset$ is in $\tilde{\mathcal{T}}$.

(T2) Suppose that $\tilde{U}$, $\tilde{V}$ are in $\tilde{\mathcal{T}}$. Then $p^{-1}(\tilde{U})$ and $p^{-1}(\tilde{V})$ are in $\mathcal{T}$ so

$$p^{-1}(\tilde{U} \cap \tilde{V}) = p^{-1}(\tilde{U}) \cap p^{-1}(\tilde{V}) \text{ is in } \mathcal{T}, \text{ so } \tilde{U} \cap \tilde{V} \text{ is in } \tilde{\mathcal{T}}.$$

(T3) Suppose that $\tilde{U}_i$ is in $\tilde{\mathcal{T}}$ for each i in some indexing set I. Then

$$p^{-1}\left(\bigcup_{i \in I} \tilde{U}_i\right) = \bigcup_{i \in I} p^{-1}(\tilde{U}_i),$$

which is in $\mathcal{T}$ since each $p^{-1}(\tilde{U}_i)$ is in $\mathcal{T}$. So $\bigcup_{i \in I} \tilde{U}_i$ is in $\tilde{\mathcal{T}}$. □

What Proposition 15.4 says is that $\tilde{U}$ is open in $X/\sim$ iff $p^{-1}(\tilde{U})$ is open in X. Note that this does *not* mean that for any open subset U of X, $p(U)$ is open in $X/\sim$; that *may* be true, but in general $p^{-1}(p(U))$ may be strictly larger than U, and may or may not be open in X (see Exercise 15.5). However, we do have

Proposition 15.5 *The natural map $p : X \to X/\sim$ is continuous when X is a space, $\sim$ is an equivalence relation on X and $X/\sim$ is given the quotient topology.*

Proof This follows since if $\tilde{U} \in \tilde{\mathcal{T}}$ then by definition of $\tilde{\mathcal{T}}$ we have $p^{-1}(\tilde{U}) \in \mathcal{T}$. □

Main property of quotients

We now touch on more general theory of quotients.

Definition 15.6 *A* quotient map *is a map $p : X \to Y$ from a space X onto a space Y such that $V \subseteq Y$ is open in Y iff $p^{-1}(V)$ is open in X.*

Notice that any quotient map is continuous from the 'only if' part of the definition.

Example 15.7 Let $p : X \to X/\sim$ be the natural map from a space X to its quotient under an equivalence relation $\sim$ as in Proposition 15.4. Then p is a quotient map.

Proof This follows from the definitions of the quotient topology and of a quotient map. □

The following main property of quotients looks rather formal, but in fact it is very useful as we shall see.

Proposition 15.8 *Suppose that $p : X \to Y$ is a quotient map and that $g : Y \to Z$ is any map to another space Z. Then g is continuous iff $g \circ p$ is continuous.*

We picture this in a diagram, Figure 15.8.

Proof If g is continuous then so is $g \circ p$ since p is continuous.

Conversely suppose that $g \circ p$ is continuous and let $U \subseteq Z$ be open in Z. Then by continuity of $g \circ p$, we have $(g \circ p)^{-1}(U)$ open in X. This says $p^{-1}(g^{-1}(U))$ is open in X, so by definition of a quotient map $g^{-1}(U)$ is open in Y. Hence g is continuous as required. □

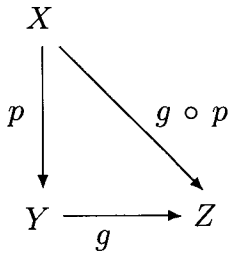

Figure 15.8. Proof of Proposition 15.8

We end this section with a useful special case of Proposition 13.15.

Proposition 15.9 *If $p : X \to Y$ is a quotient map and $A \subseteq X$ is compact then $p(A)$ is compact.*

The circle

Let X be the interval $[0, 2\pi]$ in $\mathbb{R}$, and let $\sim$ be the equivalence relation defined by: $0 \sim 2\pi$, but otherwise no two distinct points of $[0, 2\pi]$ are equivalent. Then $X/\sim$ is homeomorphic to a circle.

It is intuitively clear that the equivalence relation is designed to stick together the endpoints of the interval to construct a circle. We now prove this formally.

Proof Let S^1 denote the circle in $\mathbb{R}^2$ with centre the origin and radius 1. Define $f : [0, 2\pi] \to S^1$ by $f(t) = (\cos t, \sin t)$. From familiar properties of cos and sin we see that f is continuous, onto, and injective except that $f(0) = f(2\pi)$. So f induces a one one correspondence $g : [0, 2\pi]/\sim \to S^1$, such that $f = g \circ p$, where $p : [0, 2\pi] \to [0, 2\pi]/\sim$. Now g is continuous by Proposition 15.8. Also, $[0, 2\pi]/\sim$ is compact by Proposition 15.9, and S^1 is Hausdorff as a subspace of the metric space $\mathbb{R}^2$. Hence g is a homeomorphism by Corollary 13.27. □

This is a pattern of argument we shall use several times: to prove that a quotient space $X/\sim$ is homeomorphic to a space Y, we somehow think up a continuous map $f : X \to Y$ onto Y which 'respects the identifications', meaning that $f(x_1) = f(x_2)$ whenever $x_1 \sim x_2$, so that by Proposition 15.3 f induces a well-defined map $g : X/\sim \to Y$. Since f is onto so is g. Also, by Proposition 15.8, g is continuous. We then show that g is injective as well as onto by checking that $f(x_1) = f(x_2)$ implies $x_1 \sim x_2$. If we are lucky, we know that $X/\sim$ is compact and Y is Hausdorff, so g is a homeomorphism by Corollary 13.27.

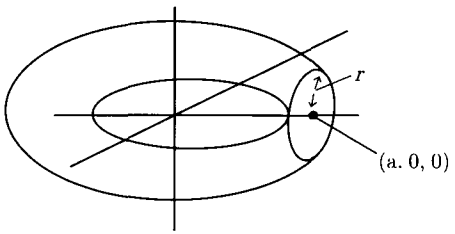

Figure 15.9. The torus in three-dimensional space

Remark In relation to the circle, we could equally well define an equivalence relation on $\mathbb{R}$ by $x \sim y$ iff x and y differ by an integer multiple of 2π, and again the quotient is homeomorphic to a circle. We follow the proof above, noting that $\mathbb{R}/\sim$ is compact since it is the continuous image of the compact space $[0, 2\pi]$ by the composition $p \circ i$ where i is the inclusion of $[0, 2\pi]$ in $\mathbb{R}$ and $p : \mathbb{R} \to \mathbb{R}/\sim$ is the natural map sending each point to its equivalence class.

As the reader may know, in this form the example is related to Fourier series: to be given a continuous 2π-periodic function on $\mathbb{R}$ is equivalent to being given a continuous function on the circle.

The torus

We now study the torus in more detail. Recall that we saw the representation of a torus as in Figure 15.2. Now our coffee-shop experience of a doughnut is not about a picture like Figure 15.2, but rather about something you can get your teeth into, as in Figure 15.9. We may describe it mathematically by taking a circle in the xz-plane and rotating it about the z-axis. If the circle has centre $(a, 0, 0)$ and radius r then we assume $r < a$ lest we get in a tangle when we rotate the circle. A general point on the circle is $(a + r\cos\theta, 0, r\sin\theta)$. When we rotate this through an angle ψ we get a general point on the torus of the form

$$((a + r\cos\theta)\cos\psi, (a + r\cos\theta)\sin\psi, r\sin\theta), \quad 0 \leqslant \theta \leqslant 2\pi, \quad 0 \leqslant \psi \leqslant 2\pi.$$

So all these points form the surface we shall call T, which looks like the surface of a doughnut. We shall prove that the space $X/\sim$ as in Figure 15.2 is homeomorphic to T.

Proof It is convenient to think of the edges of the square in Figure 15.2 as being of length 2π rather than 1. It is clear that this does not change the

homeomorphism class of the space. Let S denote the square $[0, 2\pi]\times[0, 2\pi]$ in the plane. Define $f : S \to T$ by

$$f(s, t) = ((a + r\cos t)\cos s,\ (a + r\cos t)\sin s,\ r\sin t).$$

Note that

(a) each image point is in T, so f *is* a map to T.

(b) $i \circ f : S \to \mathbb{R}^3$ is continuous since each coordinate function is continuous. Hence f is continuous by Proposition 10.6.

(c) f respects the equivalence relation $\sim$; for by periodicity of cos and sin we get $f(0, t) = f(2\pi, t)$ for any t in $[0, 2\pi]$ and $f(s, 0) = f(s, 2\pi)$ for any s in $[0, 2\pi]$; also $f(0, 0) = f(2\pi, 0) = f(0, 2\pi) = f(2\pi, 2\pi)$.

It follows from these three properties that f induces a well-defined continuous map $g : S/\sim \to T$. Since $S/\sim$ is compact by Proposition 15.9 and T is Hausdorff as a subspace of the metric space $\mathbb{R}^3$, it will follow from Corollary 13.27 that g is a homeomorphism provided we check that g is injective as well as onto. Geometrically it is clear that f is injective except for (c) above, hence g is injective. But we check this algebraically. Suppose that $f(s_1, t_1) = f(s_2, t_2)$: we want to prove $(s_1, t_1) \sim (s_2, t_2)$. We have

(i) $(a + r\cos t_1)\cos s_1 = (a + r\cos t_2)\cos s_2$;

(ii) $(a + r\cos t_1)\sin s_1 = (a + r\cos t_2)\sin s_2$;

(iii) $r\sin t_1 = r\sin t_2$.

From (iii) we get $\sin t_1 = \sin t_2$ and by considering (i)^2 + (ii)^2 we get

$$(a + r\cos t_1)^2 = (a + r\cos t_2)^2 \quad \text{so} \quad \cos t_1 = \cos t_2.$$

(Note that $a + r\cos t_1$ and $a + r\cos t_2$ are positive.) Since $t_1, t_2 \in [0, 2\pi]$, from familiar properties of cos and sin we get that either $t_1 = t_2$ or else $\{t_1, t_2\} = \{0, 2\pi\}$.

Also, from (i), (ii) and $\cos t_1 = \cos t_2$ we get $\cos s_1 = \cos s_2$ and also $\sin s_1 = \sin s_2$. So similarly either $s_1 = s_2$ or $\{s_1, s_2\} = \{0, 2\pi\}$.

In any combination of these cases we get $(s_1, t_1) \sim (s_2, t_2)$ as required. □

The real projective plane and the Klein bottle

The reader probably has an intuitive idea of what a surface is; we shall define it later. Rather than studying all surfaces systematically, we are just illustrating quotient spaces through a few standard surfaces. We obtained

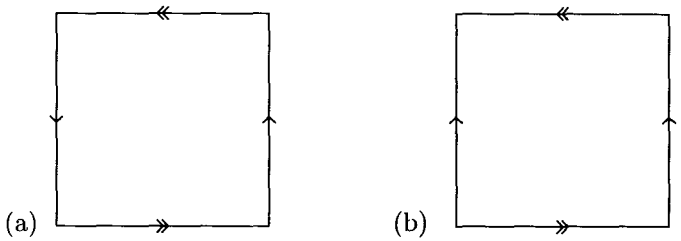

Figure 15.10. Two further quotient spaces

the torus as a quotient of a square in which opposite edges are identified (stuck together) in a certain way. As Figure 15.10 indicates, there are two more possibilities for identifying the edges of a square. The surface in Figure 15.10(a) is called *the real projective plane.* (The surface in Figure 15.10(b) is called *the Klein bottle*, and will be discussed later.) The real projective plane arises in various contexts, for example in geometry. We shall compare several different manifestations, all of them quotient spaces, and prove that the real projective plane can be embedded in $\mathbb{R}^4$; this means it is homeomorphic to a subspace of $\mathbb{R}^4$. It is an example of a *non-orientable* surface, defined later.

The real projective plane

Before describing the equivalent forms of the real projective plane, here is some notation. Let S denote the sphere in $\mathbb{R}^3$ given by the equation $x^2 + y^2 + z^2 = 1$. Let $D^+ = \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1, z \geqslant 0\}$ be the closed upper hemisphere of S. Let D be the disc in $\mathbb{R}^2$ given by $D = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant 1\}$.

Proposition 15.10 *The following quotient spaces are all homeomorphic:*

(*a*) $\mathbb{R}^3 \setminus \{0\}/\sim$ *where* $x \sim y$ *iff* $y = \lambda x$ *for some non-zero* $\lambda \in \mathbb{R}$;

(*b*) $S/\sim$ *where* $\sim$ *identifies each pair of antipodal points of* S;

(*c*) $D^+/\sim$ *where* $\sim$ *identifies each pair of antipodal points on the boundary circle of* D^+;

(*d*) $D/\sim$ *where* $\sim$ *identifies each pair of antipodal points on the boundary circle of* D;

(*e*) *the space in Figure 15.10(a).*

Each of these spaces may be called *the real projective plane*, from the viewpoint of topology. Before embarking on the proof, here are a few comments about the various forms. (There is more background on the

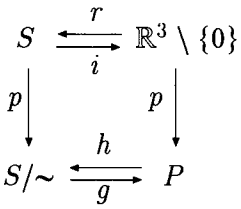

Figure 15.11. Proving part of Proposition 15.10

web site.) Modern projective geometry uses the form (a), in which we think of the projective plane as the space of lines through the origin in 3-dimensional space, and we shall therefore call it P. On the other hand Kepler's early ideas about extending plane Euclidean geometry, so that pairs of distinct lines always meet in a unique point, relate to (d). Finally we specify explicitly what $\sim$ means in (e): we begin with $[0, 1] \times [0, 1]$ and define $(s_1, t_1) \sim (s_2, t_2)$ iff one of the following holds.

(i) $s_1 = s_2$, $t_1 = t_2$;
(ii) $\{s_1, s_2\} = \{0, 1\}$ and $t_2 = 1 - t_1$;
(iii) $\{t_1, t_2\} = \{0, 1\}$ and $s_2 = 1 - s_1$.

The proof that this is an equivalence relation is on the web site.

Proof of 15.10 We prove in detail that the spaces in (a) and (b) are homeomorphic, and outline the other proofs (details are on the web site). Let us denote *any* of the quotient maps involved in this proof by p. In Figure 15.11 the map i is the inclusion of S in $\mathbb{R}^3 \setminus \{0\}$, and $r : \mathbb{R}^3 \setminus \{0\} \to S$ is given by $r(x) = \dfrac{x}{||x||}$ where $||x||$ is the length of x. It is familiar that both i and r are continuous (note that $||x|| \neq 0$ for $x \in \mathbb{R}^3 \setminus \{0\}$). Now for any $x \in S$,

$$(p \circ i)(-x) = p(-x) = p(x) = (p \circ i)(x),$$

in other words $p \circ i$ respects the identifications on S, so it induces a continuous map $g : S/\sim \to P$ as in Figure 15.11, such that $p \circ i = g \circ p$. We next show that similarly $p \circ r$ respects the identifications on $\mathbb{R}^3 \setminus \{0\}$. For any $\lambda \neq 0$ and any $x \in \mathbb{R}^3 \setminus \{0\}$ we have

$$(p \circ r)(\lambda x) = p\left(\frac{\lambda x}{||\lambda x||}\right) = p\left(\frac{\lambda}{|\lambda|}\frac{x}{||x||}\right) = p\left(\frac{\lambda}{|\lambda|} r(x)\right).$$

Now if $\lambda > 0$ then $\lambda/|\lambda| = 1$ and $p\left(\frac{\lambda}{|\lambda|} r(x)\right) = p(r(x))$, while if $\lambda < 0$

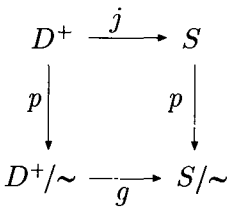

Figure 15.12. Comparing two forms of P

then $\lambda/|\lambda| = -1$ and $p\left(\frac{\lambda}{|\lambda|}r(x)\right) = p(-r(x)) = p(r(x))$. So in either case $(p \circ r)(\lambda x) = (p \circ r)(x)$, and $p \circ r$ respects the identifications on $\mathbb{R}^3 \setminus \{0\}$ as claimed. Thus $p \circ r$ induces a continuous map $h : P \to S/\sim$ as in Figure 15.11, such that $h \circ p = p \circ r$.

Finally we check that both compositions $h \circ g$ and $g \circ h$ are identity maps. First given any $x \in S$ we have $r(i(x)) = r(x) = x/||x|| = x$ since $||x|| = 1$, so $r \circ i$ is the identity map of S, and

$$(h \circ g) \circ p = h \circ (g \circ p) = h \circ (p \circ i) = (h \circ p) \circ i = (p \circ r) \circ i = p \circ (r \circ i) = p.$$

Now $p : S \to S/\sim$ is onto, so $h \circ g$ is the identity map of $S/\sim$. Similarly for any $x \in \mathbb{R}^3 \setminus \{0\}$ we have $i(r(x)) = i(x/||x||) = x/||x||$, so

$$(g \circ h)(p(x)) = g((h \circ p)(x)) = g((p \circ r)(x)) = (g \circ p)(r(x)) =$$
$$(p \circ i)(r(x)) = p((i \circ r)(x)) = p\left(\frac{x}{||x||}\right) = p(x),$$

where the last equality follows since $1/||x|| \neq 0$. Now $p : \mathbb{R}^3 \setminus \{0\} \to P$ is onto so $g \circ f$ is the identity map of P. This proves that P and $S/\sim$ are homeomorphic.

For the rest of the proof it is useful to know that P is Hausdorff. We could give a direct proof of this, but it also has an indirect proof: from above there is a homeomorphism between P and $S/\sim$; Proposition 15.11 below gives a continuous injective map from $S/\sim$ into $\mathbb{R}^4$; so P is Hausdorff since $\mathbb{R}^4$ is (Proposition 11.7(c)).

We now outline the proof that the spaces in (b) and (c) are homeomorphic. We use Figure 15.12, which is similar to Figure 15.11. In Figure 15.12 the map j is inclusion. We may check that $p \circ j$ respects the identifications on D^+ so induces a continuous map $g : D^+/\sim \to S/\sim$. It can be checked that g is a one-one correspondence. But $D^+/\sim$ is compact by Proposition 15.9 and $S/\sim$ is Hausdorff, so g is a homeomorphism by Corollary 13.27.

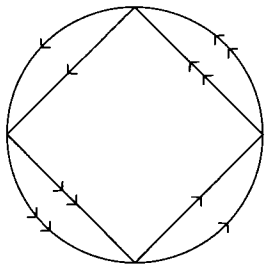

Figure 15.13. Square inside disc

An informal way of proving that the spaces in (c) and (d) are homeomorphic is to say 'just sit on D^+ and flatten it down to D' (the identifications on the boundary are unchanged). The formal version is on the web site.

Finally, to see that the spaces in (d) and (e) are homeomorphic, we may fit a square inside D as in Figure 15.13. Now map the square radially outwards onto the disc. The identifications on the boundary of the square correspond to those on the boundary of D shown in Figure 15.13. But a little thought shows that these simply mean that antipodal points on the boundary of D are identified. □

The last thing we prove about the real projective plane is that it is homeomorphic to a subspace of $\mathbb{R}^4$; we say 'the real projective plane can be embedded in $\mathbb{R}^4$'.

Proposition 15.11 *There is a homeomorphism from P to a subspace of $\mathbb{R}^4$.*

Proof In view of Proposition 15.10 we can operate with any of the forms of P listed there. We choose $S/\sim$. Define a map $f : S \to \mathbb{R}^4$ by

$$f(x, y, z) = (x^2 - y^2, xy, yz, zx).$$

Since each coordinate function is continuous, this is a continuous map into $\mathbb{R}^4$. It respects the identifications on S: for $(x_1, y_1, z_1) \sim (x_2, y_2, z_2)$ implies that $(x_1, y_1, z_1) = \pm(x_2, y_2, z_2)$, giving $f(x_1, y_1, z_1) = f(x_2, y_2, z_2)$. Hence by Proposition 15.8, f induces a continuous map $g : S/\sim \to \mathbb{R}^4$ such that $f = g \circ p$ where $p : S \to S/\sim$ is the natural map. Since $S/\sim$ is compact by Proposition 15.9 and $\mathbb{R}^4$ is Hausdorff, it is now enough by Corollary 13.27 to prove that g is injective. We recall that for this it is enough to prove that $f(x_1, y_1, z_1) = f(x_2, y_2, z_2)$ implies $(x_1, y_1, z_1) \sim (x_2, y_2, z_2)$, for points (x_1, y_1, z_1), (x_2, y_2, z_2) in S. The algebra needed to do this is on the web site. □

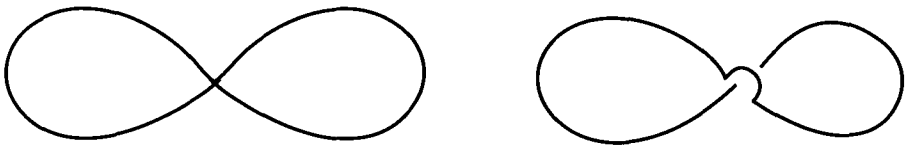

Figure 15.14. Figure-of-eight deformed to embedding

The Klein bottle

The final surface we consider in detail is the Klein bottle, which is represented in Figure 15.10(b). We find it convenient to change the scale, not affecting the topological type.

Definition 15.12 *The* Klein bottle K *is the quotient of* $[0, 2\pi] \times [0, \pi]$ *by the equivalence relation defined by:* $(x_1, y_1) \sim (x_2, y_2)$ *iff one of the following holds*

(i) $x_1 = x_2$ *and* $y_1 = y_2$;
(ii) $\{y_1, y_2\} = \{0, \pi\}$ *and* $x_2 = 2\pi - x_1$;
(iii) $\{x_1, x_2\} = \{0, 2\pi\}$ *and* $y_1 = y_2$;
(iv) $x_1, x_2 \in \{0, 2\pi\}$ *and* $y_1, y_2 \in \{0, \pi\}$.

The proof that $\sim$ is an equivalence relation is on the web site. It is strongly recommended that you type 'Klein bottle' into a search engine on the internet; there are some good pictures, for example of a glass model of the Klein bottle immersed in Euclidean three-space. 'Immersed' means roughly that it is embedded except that it intersects itself in a manner that is not too wild. You can perhaps persuade yourself that the next proposition is true as follows. A figure-of-eight in the plane is an immersion of a circle, and it can be deformed slightly to become an embedding of a circle in three-space by taking a small interval around the self-intersection point in one of the intersecting pieces of the circle, and moving it into the third dimension to become a small arc which 'misses' the other piece of the circle (Figure 15.14). It is plausible that one can likewise get rid of the self-intersection circle of a Klein bottle in three-space (Figure 15.15) by deforming slightly in the fourth dimension so that the bottle 'misses' itself in $\mathbb{R}^4$.

Proposition 15.13 *The Klein bottle can be embedded in* $\mathbb{R}^4$.

Proof Write $X = [0, 2\pi] \times [0, \pi]$. Define $f : X \to \mathbb{R}^4$ as follows: for $(x, y) \in X$, let

$$f(x, y) = ((2 + \cos x) \cos 2y, (2 + \cos x) \sin 2y, \sin x \cos y, \sin x \sin y).$$

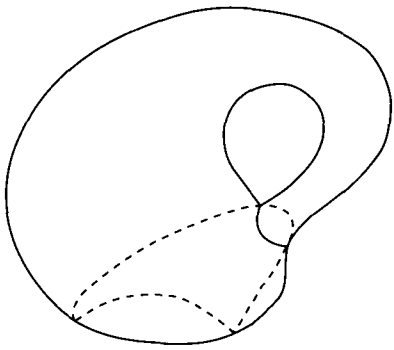

Figure 15.15. Immersion of the Klein bottle in three-space

Since each of its coordinate functions is continuous, f is continuous. Also

$$f(2\pi - x, \pi) = (2 + \cos x, 0, (-\sin x)(-1), 0) = f(x, 0).$$

Similarly

$$f(2\pi, y) = (3\cos 2y, 3\sin 2y, 0, 0) = f(0, y).$$

From this it follows that f respects the identifications on X and therefore induces a continuous map $g : X/\sim \to \mathbb{R}^4$. Since K is compact (by Proposition 15.9) and $\mathbb{R}^4$ is Hausdorff, it remains to show that g is injective, or $f(x_1, y_1) = f(x_2, y_2)$ implies $(x_1, y_1) \sim (x_2, y_2)$ for (x_1, y_1), $(x_2, y_2) \in X$. Now $f(x_1, y_1) = f(x_2, y_2)$ gives

$$\text{(i)} \quad (2 + \cos x_1)\cos 2y_1 = (2 + \cos x_2)\cos 2y_2;$$

$$\text{(ii)} \quad (2 + \cos x_1)\sin 2y_1 = (2 + \cos x_2)\sin 2y_2;$$

$$\text{(iii)} \quad \sin x_1 \cos y_1 = \sin x_2 \cos y_2;$$

$$\text{(iv)} \quad \sin x_1 \sin y_1 = \sin x_2 \sin y_2.$$

Taking (i)^2+(ii)^2 gives $(2 + \cos x_1)^2 = (2 + \cos x_2)^2$ so $\cos x_1 = \cos x_2$. Now using (i) and (ii) again we get $\cos 2y_1 = \cos 2y_2$, $\sin 2y_1 = \sin 2y_2$. So *either* $y_1 = y_2$ *or* $\{2y_1, 2y_2\} = \{0, 2\pi\}$, i.e. $\{y_1, y_2\} = \{0, \pi\}$.

Case 1: Suppose $y_1 = y_2$. Then (iii) and (iv) give $\sin x_1 = \sin x_2$. Since we saw earlier that $\cos x_1 = \cos x_2$, we get that *either* $x_1 = x_2$ *or else* $\{x_1, x_2\} = \{0, 2\pi\}$.

Case 2: Suppose $\{y_1, y_2\} = \{0, \pi\}$. Then (iii) gives $\sin x_1 = -\sin x_2$, and since $\cos x_1 = \cos x_2$ we get $x_2 = 2\pi - x_1$.

We can check that in any combination of these cases, $(x_1, y_1) \sim (x_2, y_2)$. □

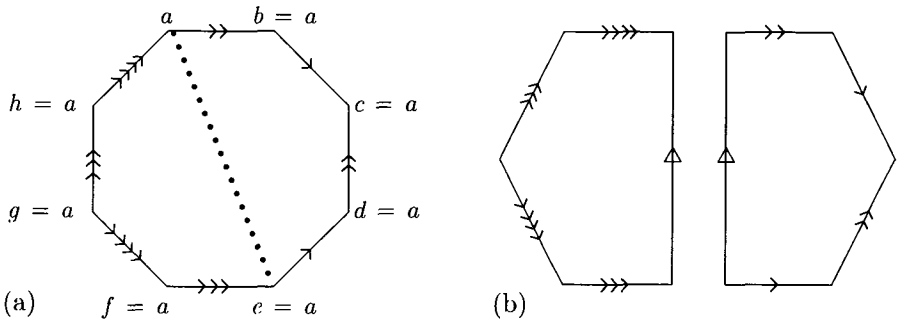

Figure 15.16. (a) Octagon (edges stuck together) and (b) octagon cut (and deformed)

Remark In view of Propositions 15.11 and 15.13 the reader may suspect, or may already know, that neither the real projective plane nor the Klein bottle can be embedded in $\mathbb{R}^3$. A proof of this which is both rigorous and accessible is difficult to find. There is a heuristic (which means non-rigorous) argument for the real projective plane on the web site.

Cutting and pasting

Further geometric motivation for quotient spaces is afforded by a technique called 'cutting and pasting'. The idea is that we can cut a space into several pieces and as long as we join the cut edges together again in the correct direction at some stage, it is intuitively clear that the topological type of the object is not affected.

Suppose an octagon has its edges stuck together as in Figure 15.16(a). As it stands, it is a little difficult to visualize what this gives. But suppose we cut it along the straight line ae, and consider the two pieces separately, as in Figure 15.16(b), with arrows on the two edges of the cut to show that they should be stuck together at some stage. We know from earlier that each of these pieces on its own is a torus with a hole in it, and the boundaries of the holes are supposed to be stuck together. It is geometrically clear that what we get finally is a double torus, or surface of a kind of pretzel—see Figure 15.17.

For another example, suppose we have a hexagon with edges stuck together in the way that the arrows indicate in Figure 15.18. Again it is not immediately obvious what this gives. But suppose that we cut it into three pieces along the dashed lines, and as in Figure 15.19 put arrows on the cut edges to show how they should be assembled.

Now let us reassemble it in stages. First put the two triangles together to get Figure 15.20.

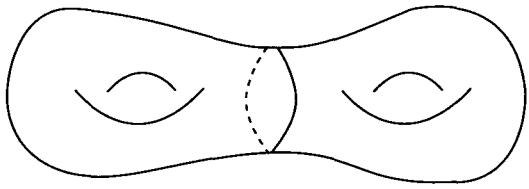

Figure 15.17. Double torus

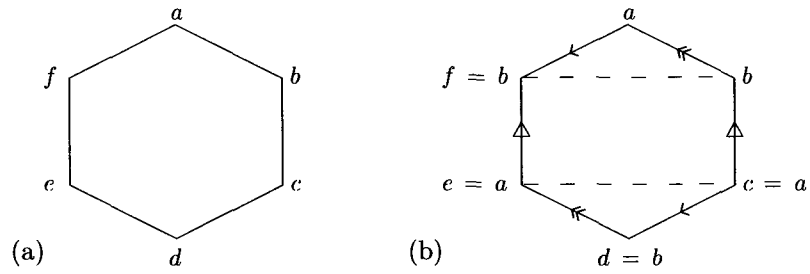

Figure 15.18. (a) Hexagon and (b) hexagon with edges stuck together

Now straighten up the parallelogram and stick it to the other piece as in Figure 15.21. We see that the pasting instructions are equivalent to those for a torus in Figure 15.2, so what we get is a torus.

The shape of things to come

We have provided only a taster of surfaces above. There are more general approaches, and one can classify 'all' surfaces up to homeomorphism. A more modest goal would be to classify all 'closed' surfaces. A general definition of a closed surface is: a compact Hausdorff space each point of which is contained in an open set which is homeomorphic to an open disc in the plane. So a closed surface is *locally Euclidean.* The need to include

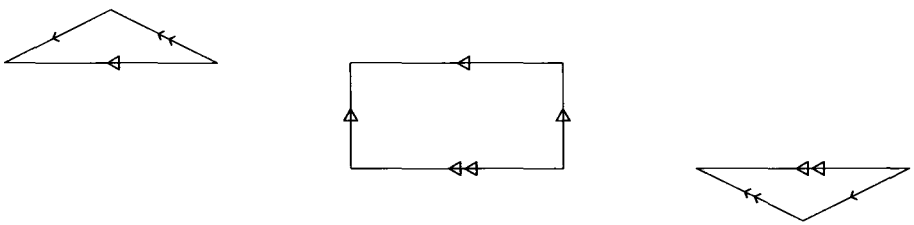

Figure 15.19. The hexagon cut up

Figure 15.20. Hexagon partially reassembled

Figure 15.21. Surface reassembled—two equivalent forms

'Hausdorff' here may seem surprising; the web site has a pathological example to illustrate why we need this condition.

Closed surfaces naturally divide into *orientable* and *non-orientable* ones. The quickest way to explain this division is to say: a surface is non-orientable iff it contains a Möbius band. For example, Figure 15.22 illustrates that the Klein bottle is non-orientable.

We look briefly at one way of constructing closed surfaces, again using quotient spaces.

(1) Attaching handles to a sphere: take a sphere S^2 with two open discs removed and glue in a cylinder to make a surface homeomorphic to a torus (Figure 15.23). Explicitly, let $(S^2 \setminus (D_1 \cup D_2)) \sqcup (S^1 \times I)$ be the disjoint union of a 2-sphere with two open discs removed, and the cylinder given by the product of a circle S^1 with a closed unit interval I. We then form

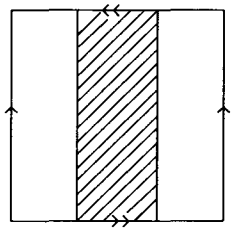

Figure 15.22. A Möbius band in a Klein bottle

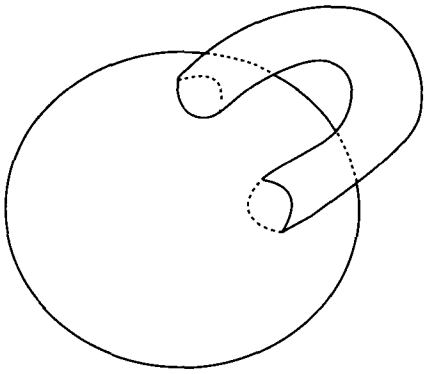

Figure 15.23. Adding a handle to a sphere

the quotient space by the equivalence relation telling you how to stick the ends of the cylinder to the boundaries of the holes in the sphere. More generally, we can attach g such handles in an orderly fashion and get what is called an orientable surface of genus g.

(2) Sewing a Möbius band into a sphere: take a sphere with just one hole in it, and attach a Möbius band by sticking its boundary circle onto the boundary of the hole. What do we get? Cutting and pasting will tell. First, the sphere with an open disc removed is topologically the same as a closed disc. So what we get is like sewing the boundary of a disc to the free edge of a Möbius band. In Figure 15.24 we first cut the disc in half then sew on the semi-circular parts of their boundaries as the letters indicate. What we get is homeomorphic to a real projective plane. Again, more generally we can sew in g Möbius bands in an orderly fashion.

The (topological) classification of closed surfaces says: any orientable closed surface is homeomorphic to a sphere with g handles for some integer $g \geqslant 0$, while any non-orientable closed surface is homeomorphic to a

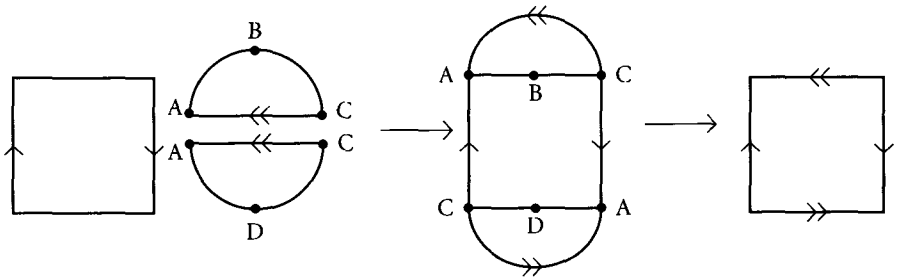

Figure 15.24. A sphere with a Möbius band sewn in

sphere with g Möbius bands attached for some integer $g > 0$. There is actually more to the result than this: it tells you how to find out what g is for a given surface in terms of a number called its *Euler characteristic.* For further information we refer for example to Lawson (2003).

Exercise 15.1 In this question a triangle includes the interior as well as the edges, and identifying line segments ab, cd means that each point on ab is identified with its image under the linear map from ab onto cd which maps a to c and b to d.

Without rigorous reasons, explain geometrically what you get if

(a) in the real interval $[0, 2]$ you identify the points 0, 1, and 2;

(b) in the disjoint union of two triangles abc, xyz, you identify the pairs of edges ab, xy; bc, yz; ca, zx; (so ab is identified with xy and so on).

(c) in a triangle abc, you identify the edges ca, ab;

[*Hint: for (c) try 'cutting and pasting', cutting from a to the mid-point of bc.*]

Exercise 15.2 Recall that for a Möbius band we let $X = [0, 1] \times [0, 1]$ and for (s_1, t_1), $(s_2, t_2) \in X$, we let $(s_1, t_1) \sim (s_2, t_2)$ iff one of the following holds:

(i) $s_1 = s_2$ and $t_1 = t_2$;

(ii) $s_1 = 0$. $s_2 = 1$ and $t_2 = 1 - t_1$;

(iii) $s_1 = 1$, $s_2 = 0$ and $t_2 = 1 - t_1$

Show that this relation is reflexive and transitive.

Exercise 15.3 For the equivalence relation in Exercise 15.2 show that the equivalence classes are as follows: if $0 < s < 1$ then the singleton set $\{(s.\ t)\}$ is an equivalence class on its own for any fixed $t \in [0, 1]$, while for each $t \in [0, 1]$ the pair $\{(0, t), (1, 1 - t)\}$ is an equivalence class.

Exercise 15.4 Let M be the Möbius band obtained from $[0, 1] \times [0, 1]$ by identifying the points $(0. y)$ and $(1. 1 - y)$ for each $y \in [0, 1]$. Show that cutting along the image in M of the line segment joining the points $(0, 1/2)$ and $(1, 1/2)$ produces a space homeomorphic to $S^1 \times [0, 1)$. Show also that cutting along the images in M of the line segment from $(0, 1/3)$ to $(1, 1/3)$ and the line segment from $(0, 2/3)$ to $(1, 2/3)$ produces the disjoint union of an 'open' Möbius band (identification space of $[0, 1] \times (0, 1)$) and a cylinder (homeomorphic to $[0, 1] \times (0, 1]$).

Exercise 15.5 Let $f : [0, 2\pi] \to S^1$ be the map defined by $f(t) = (\cos t, \sin t)$. Show that $[0, \pi)$ is open in $[0, 2\pi]$ but $f([0, \pi))$ is not open in S^1.

Exercise 15.6 Define an equivalence relation on $[0, 1]$ by: $x \sim y$ if and only if either both of x, y are rational or both are irrational. Check that this is an equivalence relation, and prove that the corresponding quotient space is the

two-point space with the indiscrete topology. (Note that $[0, 1]$ is Hausdorff but $[0, 1]/\sim$ is not Hausdorff.)

Exercise 15.7 Prove that the following is a necessary and sufficient condition for a map $f : X \to Y$ from a space X onto a space Y to be a quotient map: a subset V of Y is closed in Y iff $f^{-1}(V)$ is closed in X.
[Hint: use Corollary 3.8.]

Exercise 15.8 Prove that the composition of two quotient maps is again a quotient map.

Exercise 15.9 Define an equivalence relation on $\mathbb{R}^2$ by $(x, y) \sim (x', y')$ iff $x - x'$ and $y - y'$ are both integral mutiples of 2π. Prove that the quotient space $\mathbb{R}^2/\sim$ is homeomorphic to the torus $T \subseteq \mathbb{R}^3$ defined by

$$T = \{((a + r\cos\theta)\cos\psi,\ (a + r\cos\theta)\sin\psi,\ r\sin\theta); 0 \leqslant \psi \leqslant 2\pi,\ 0 \leqslant \theta \leqslant 2\pi\},$$

where $0 < r < a$.

16 Uniform convergence

Motivation

We now move towards the third of a trio of important concepts: connectedness, compactness and completeness. We shall study completeness in the context of metric spaces. An important ingredient in establishing completeness for several of our metric spaces will be uniform convergence. In this chapter, we begin the study of uniform convergence in an elementary fashion; later in the chapter we express it in terms of convergence in function spaces.

The study of uniform convergence may be motivated as follows. Many particular functions in analysis are studied by means of sequences or series. Likewise existence proofs for solutions of differential and other equations often produce the solution as the limit of a sequence of functions. In such cases, we wish to know whether the limit is continuous. (Similarly, although we do not study this here, we want to know whether we can differentiate or integrate the limit by differentiating or integrating each term in the sequence and taking the limit.) Uniform convergence contributes to sufficient, though not usually necessary, conditions for ensuring that the limit function is well-behaved if the terms in the sequence are well-behaved.

Definition and examples

Initially we shall be concerned with real-valued functions defined on some subset $D \subseteq \mathbb{R}$, usually an interval. Suppose that (f_n) is a sequence of real-valued functions, and that the domain of each contains D. If we fix attention on a particular point $x \in D$, the values of the functions f_n at x give a sequence $(f_n(x))$ in $\mathbb{R}$. Suppose that for each $x \in D$ the sequence $(f_n(x))$ converges. Then we may define a new function $f : D \to \mathbb{R}$ by putting $f(x) = \lim_{n\to\infty} f_n(x)$.

A slightly different viewpoint is to suppose given a sequence of real-valued functions $\{f_n : n \in \mathbb{R}\}$ and a real-valued function f, all with domain D, and to make the following definition.

Definition 16.1 *The sequence* (f_n) converges to f pointwise on D *if for each* $x \in D$ *the real number sequence* $(f_n(x))$ *converges to* $f(x)$.

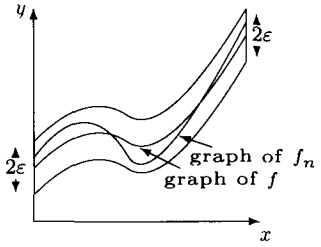

Figure 16.1. Uniform convergence

Pointwise convergence is sometimes not very well-behaved; for example we cannot deduce continuity of f from continuity of all the f_n.

Example 16.2 Let $f_n(x) = x^n$ for all $x \in [0, 1]$ and $f(x) = \lim_{n\to\infty} f_n(x)$. Then

$$f(x) = \begin{cases} 0, & x \in [0, 1), \\ 1, & x = 1. \end{cases}$$

Thus the pointwise limit function is discontinuous at 1 although every f_n is continuous. (The reader can gain insight here by drawing the graphs of f_n for $n = 1, 2, 3, 4$.)

In certain situations (e.g. in Fourier series) this kind of thing must happen. But often we should like the limit function to be continuous. Uniform convergence helps to ensure this.

Definition 16.3 *A sequence (f_n) of real-valued functions defined on a domain $D \subseteq \mathbb{R}$ converges to a function f* uniformly on D *if given $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \varepsilon$ for all $n \geqslant N$ and all $x \in D$.*

If this is fulfilled we write '$f_n \to f$ uniformly on D' and call f 'the uniform limit of f_n'. Clearly if $f_n \to f$ uniformly on D then $f_n \to f$ pointwise on D. Uniform convergence is stronger in that given $\varepsilon > 0$ there must exist an integer N which does the necessary job for all x in D simultaneously, while for pointwise convergence, given $\varepsilon > 0$ we may use a different N_x for each x in D. Thus uniform convergence is global in D.

Here is a graphical interpretation of uniform convergence. Suppose $D = [0, 1]$ and let us draw the graph of f (Figure 16.1). Uniform convergence of (f_n) to f on D means that if we draw the ribbon of vertical width $2\varepsilon > 0$ centred on the graph of f, then no matter how small ε is,

there is some stage in the sequence beyond which the graph of every f_n lies within this ribbon.

Remarks (a) Suppose that a sequence (a_n) of real numbers converges to a. For any $D \subseteq \mathbb{R}$, let $f_n : D \to \mathbb{R}$, $f : D \to \mathbb{R}$ be the constant functions, $f_n(x) = a_n$, $f(x) = a$ for all $x \in D$. Then it is easy to see that (f_n) converges to f uniformly on D.
(b) If (f_n) converges to f uniformly on D and $D' \subseteq D$ then clearly (f_n) converges to f uniformly on D'.

We now show that convergence in Example 16.2 is non-uniform. To do so, we put the negative of Definition 16.3 into the following usable form: there exists some $\varepsilon_0 > 0$ such that for any $N \in \mathbb{N}$ there is an $x \in D$ and an $n \geqslant N$ such that $|f_n(x) - f(x)| \geqslant \varepsilon_0$. In Example 16.2 we may take $\varepsilon_0 = 1/2$. The idea now is to show that no matter how large the integer N is, we may choose an $x \in [0, 1]$, close to but not equal to 1, such that $x^N \geqslant 1/2$. For such x we have $|f_N(x) - f(x)| = |x^N - 0| = x^N \geqslant 1/2$, so convergence is not uniform on $[0, 1]$. If we take $x = 2^{-1/N}$ then $x^N = 1/2$ as required.

Warning The same proof shows that convergence of (x^n) to the zero function is not uniform on $[0, 1)$ either. It is tempting to think that by dropping one 'bad' point the convergence will become uniform.

On the other hand, if we take $D = [0, a]$ for some $a \in (0, 1)$ then it is true that $x^n \to 0$ uniformly on D as $n \to \infty$. For given $\varepsilon > 0$ we may choose N large enough so that $a^N < \varepsilon$. Then for any $n \geqslant N$ and any $x \in [0, a]$ we have $|x^n - 0| \leqslant a^n \leqslant a^N < \varepsilon$.

This success may be explained as follows. For fixed $a \in (0, 1)$ we were able to locate an upper bound a^n for $|f_n(x) - f(x)|$ on $[0, a]$, and then we used the fact that $a^n \to 0$ as $n \to \infty$. This suggests a criterion for uniform convergence, which is really just a translation of the definition.

Proposition 16.4 *Let $f, f_n : D \to \mathbb{R}$ be real-valued functions on D. Then $f_n \to f$ uniformly on D if $M_n = \sup_{x \in D} |f_n(x) - f(x)|$ exists for all sufficiently large n and $M_n \to 0$ as $n \to \infty$.*

Proof Suppose that $M_n \to 0$ as $n \to \infty$. Given $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $0 \leqslant M_n < \varepsilon$ whenever $n \geqslant N$. Since by definition of M_n we have $|f_n(x) - f(x)| \leqslant M_n$ for all $x \in D$, we get $|f_n(x) - f(x)| < \varepsilon$ for all $n \geqslant N$ and all $x \in D$.

Conversely suppose that $f_n \to f$ uniformly on D as $n \to \infty$. Then given $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \varepsilon/2$ for all

$n \geqslant N$ and all $x \in D$. Now the set $S_n = \{|f_n(x) - f(x)| : x \in D\}$ is bounded above by $\varepsilon/2$, so $M_n = \sup_{x \in D} S_n \leqslant \varepsilon/2 < \varepsilon$ for all $n \geqslant N$ as required. □

Along with elementary calculus, this criterion often allows us to determine whether particular sequences are uniformly convergent. We give positive and negative examples of this.

Example 16.5 Let $f_n(x) = n^2x(1-x)^n$ for $x \in [0, 1]$. We check first that (f_n) converges pointwise to the zero function on $[0, 1]$. For if $x = 0$ or 1 then $f_n(x) = 0$ for all n so certainly $(f_n(x))$ converges to 0. If $x \in (0, 1)$, then $0 < f_n(x) < n^2(1 - x)^n$ so $(f_n(x))$ converges to 0 by Exercise 4.9. Let us test for uniform convergence. To find

$$M_n = \sup_{x \in [0,1]} |f_n(x) - f(x)| = \sup_{x \in [0,1]} f_n(x)$$

we use calculus. We calculate $f_n'(x) = n^2((1-x)^{n-1}[1-(n+1)x])$. Since $f_n(0) = 0 = f_n(1)$, while $f_n(x) \geqslant 0$ for $x \in [0, 1]$, the maximum of f_n in $[0, 1]$ is attained where $f_n'(x) = 0$ in $(0, 1)$, namely at $x = 1/(n+1)$. This tells us that

$$M_n = \frac{n^2 n^n}{(n+1)^{n+1}} = \frac{n^{n+2}}{(n+1)^{n+1}} \to \infty \quad \text{as} \quad n \to \infty.$$

Hence by Proposition 16.4 convergence is not uniform on $[0, 1]$.

Example 16.6 Let $f_n(x) = xe^{-nx^2}$ for $x \in [0, 1]$. Again we may show that (f_n) converges pointwise to the zero function on $[0, 1]$. Also, working as in Example 16.5 we may show that M_n is attained at $x = (2n)^{-1/2}$ and takes value $(2en)^{-1/2} \to 0$ as $n \to \infty$. Thus convergence is uniform on $[0, 1]$.

Example 16.7 Finally we give an example in geometric language, featuring the functions f_n defined in Example 14.23. (It is suggested that you look back at Figure 14.1 to follow this.) Then (f_n) converges pointwise on $[0, 1]$ to the zero function. First, $f_n(0) = 0$ for all n. Also, if $x > 0$ then for all n sufficiently large 'all the action has passed to the left of x' so $f_n(x) = 0$. Explicitly, whenever $1/2^{n-1} < x$, we have $f_n(x) = 0$. But is this convergence uniform? We can see geometrically that it is not, by thinking about a ribbon of vertical width 1 centred on the zero function, so having its upper boundary at height $1/2$. Then no matter how large n is, the graph of f_n will have a spike sticking out above the ribbon. Analytically, in this case $M_n = 1$ for all n, and $M_n \not\to 0$ as $n \to \infty$.

Proposition 16.4 also has a theoretical application: we can use it to reinterpret uniform convergence as (ordinary) convergence in a function space with the sup metric. Let $(\mathcal{B}(D, \mathbb{R}), d_\infty)$ be the metric space of all bounded real-valued functions on D, with the sup metric

$$d_\infty(f, g) = \sup_{x \in D} |f(x) - g(x)|.$$

It follows from Definition 6.25 that if $f_n \to f$ in the space $(\mathcal{B}(D, \mathbb{R}), d_\infty)$ then $d_\infty(f_n, f) \to 0$ as $n \to \infty$. But by Proposition 16.4 this means that $f_n \to f$ uniformly on D as $n \to \infty$.

We can also talk about uniform convergence of functions that are not necessarily bounded: for example $(1/x + 1/n)$ converges to $1/x$ uniformly on $(0, 1)$, although every function involved is unbounded on $(0, 1)$. In fact this *can* still be interpreted as convergence in a suitable function space, but we shall restrict attention to bounded functions in our function spaces.

Cauchy's criterion

Recall that Cauchy sequences were important in the section of Chapter 4 on sequences of real numbers. Just as there is a concept of uniform convergence, so too there is a concept of being uniformly Cauchy.

Definition 16.8 *A sequence (f_n) of real-valued functions defined on a domain $D \subseteq \mathbb{R}$ is said to be* uniformly Cauchy on D *if given $\varepsilon > 0$ there exists an integer N such that $|f_m(x) - f_n(x)| < \varepsilon$ for all $m, n \geqslant N$ and all $x \in D$.*

As usual, the 'uniform' feature is that the same N has to work for all $x \in D$, so the condition is global in D. Again here we emphasize that we are restricting to the case when all the functions concerned are bounded. In this case, as for convergence, a sequence (f_n) as above is uniformly Cauchy iff it is Cauchy as a sequence in $(\mathcal{B}(D, \mathbb{R}), d_\infty)$ in the sense of Definition 6.27.

Theorem 16.9 (Cauchy's criterion for uniform convergence.) *Let (f_n) be a sequence of real-valued functions defined on $D \subseteq \mathbb{R}$. Then (f_n) converges uniformly on D iff it is uniformly Cauchy on D.*

Proof The 'only if' part is similar to the corresponding part of Theorem 4.18 and is omitted.

Suppose that (f_n) is uniformly Cauchy on D. For each $x \in D$ the real sequence $(f_n(x))$ is Cauchy and hence, using completeness of $\mathbb{R}$, $(f_n(x))$

converges to a real number which we label $f(x)$. This determines a real-valued function f on D. Our goal is to show that (f_n) converges to f uniformly in D.

Given $\varepsilon > 0$ there exists an integer N such that $|f_m(x) - f_n(x)| < \varepsilon$ for all $m, n \geqslant N$ and all $x \in D$. There are two ways of stating the rest of the proof.

(a) For each $x \in D$, convergence of $(f_n(x))$ to $f(x)$ tells us there is an integer $n(x)$ such that $|f_{n(x)}(x) - f(x)| < \varepsilon$, and moreover we may choose $n(x) \geqslant N$. Hence for any $m \geqslant N$ and all $x \in D$ we have

$$|f_m(x) - f(x)| \leqslant |f_m(x) - f_{n(x)}(x)| + |f_{n(x)}(x) - f(x)| < 2\varepsilon.$$

(Note that N is independent of $x \in D$, although $n(x)$ is not.)

(b) In the equality $|f_m(x) - f_n(x)| < \varepsilon$ for $m, n \geqslant N$ and all $x \in D$ we may keep m fixed and let $n \to \infty$, to get $|f_m(x) - f(x)| \leqslant \varepsilon$ for all $m \geqslant N$ and all $x \in D$.

Either way, (f_n) converges uniformly on D. □

Remarks (a) Cauchy's criterion for uniform convergence has the advantage that the limit function need not be known in advance in order to prove uniform convergence. This is particularly useful when we are trying to define a function as the limit of a sequence of functions.

(b) When all the functions involved are bounded, Cauchy's criterion says that the metric space $(\mathcal{B}(D, \mathbb{R}), d_\infty)$ is like the real numbers – any Cauchy sequence in it converges. We study such spaces further in Chapter 17.

(c) As for sequences in $\mathbb{R}$ we can translate the above results into the language of series (of functions), and develop tests for uniform convergence of such series.

Uniform limits of sequences

Theorem 16.10 *If $f_n : (a, b) \to \mathbb{R}$ is continuous at $c \in (a, b)$ for every $n \in \mathbb{N}$ and if $f_n \to f$ uniformly on (a, b) then f is continuous at c.*

Proof This is a 3ε-argument. Given $\varepsilon > 0$, by uniform convergence there exists $N \in \mathbb{N}$ such that $|f_n(x) - f(x)| < \varepsilon$ for all $n \geqslant N$ and all $x \in (a, b)$. Now f_N is continuous at c so there exists $\delta > 0$ such that $|f_N(x) - f_N(c)| < \varepsilon$ for all $x \in (a, b)$ with $|x - c| < \delta$. Hence for

any such x,

$$|f(x) - f(c)| \leqslant |f(x) - f_N(x)| + |f_N(x) - f_N(c)| + |f_N(c) - f(c)| < 3\varepsilon,$$

which shows that f is continuous at c. □

Remarks (a) Theorem 16.10 shows how uniform convergence can sometimes validate the interchange of two limiting processes. We shall see that it says

$$\lim_{x\to c} \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \lim_{x\to c} f_n(x).$$

For the inner limit on the left-hand side is $f(x)$ by convergence of $(f_n(x))$ to $f(x)$, and the inner limit on the right-hand side is $f_n(c)$ by continuity of f_n at c. So the whole equation says that the limit as $x \to c$ of $f(x)$ is the limit as $n \to \infty$ of $f_n(c)$, namely $f(c)$. This says f is continuous at c.

(b) In terms of function spaces of bounded functions, Theorem 16.10 says that in $\mathcal{B} = (\mathcal{B}((a, b), \mathbb{R}), d_\infty)$, the subspace $\mathcal{C}_c$, of those functions which are continuous at c, is closed in $\mathcal{B}$. For if $f \in \overline{\mathcal{C}_c}$ then by Exercise 6.26 there is a sequence (f_n) in $\mathcal{C}_c$ converging to f (in the sup metric), and by Theorem 16.10 $f \in \mathcal{C}_c$. So $\mathcal{C}_c$ is closed in $\mathcal{B}$. Since any intersection of closed sets is closed, it follows that for any subset $C \subseteq (a, b)$ the subspace $\mathcal{C}_C$ of $\mathcal{B}$ consisting of those functions which are continuous on C, is closed in $\mathcal{B}$.

Theorem 16.10 may be a little austere, but it has useful consequences. The one-sided analogue works similarly, and using this as well we get

Corollary 16.11 *Suppose for each $n \in \mathbb{N}$ the function $f_n : [a, b] \to \mathbb{R}$ is continuous, and that (f_n) converges to a function f uniformly on $[a, b]$. Then f is continuous on $[a, b]$.*

One practical outcome of this is an easy way of detecting non-uniform convergence, when it works.

Corollary 16.12 *Suppose that the pointwise limit of a sequence (f_n) of continuous functions on $[a, b]$ is not continuous on $[a, b]$. Then the convergence is not uniform.*

Example 16.13 Let $f_n(x) = 1/(1 + nx)$ for $x \in [0, 1]$. Then f_n is continuous, and if the pointwise limit of (f_n) is written as f then

$$f(x) = \begin{cases} 0 & x \in (0, 1] \\ 1 & x = 0. \end{cases}$$

The fact that the pointwise limit is discontinuous shows that the convergence is not uniform.

Generalizations

★ Uniform convergence makes sense for sequences of maps from any set D to any metric space X. In particular, uniform convergence for sequences of bounded maps from D to X can be expressed as ordinary convergence in the metric space $\mathcal{B}(D, X)$ with the sup metric. If D is a topological space and $c \in D$ we may consider the subspace $\mathcal{C}_c \subseteq \mathcal{B}(D, X)$ of those maps which are continuous at c. The proof of Theorem 16.10 readily extends to show that $\mathcal{C}_c$ is closed in $\mathcal{B}(D, X)$. The same result holds when c is replaced by a general subspace C of D. ★

Exercise 16.1 Suppose that $D \subseteq \mathbb{R}$ is of the form $D = \bigcup_{i=1}^{r} D_i$ and that a sequence of functions $f_n : D \to \mathbb{R}$ is uniformly convergent on D_i for each $i = 1, 2, \ldots, r$. Show that (f_n) is uniformly convergent on D.

In Exercises 16.2, 16.4, 16.5, 16.8 suppose that $D \subseteq \mathbb{R}$ and that f_n, g_n are real-valued functions defined on D.

Exercise 16.2 Given that (f_n), (g_n) converge to functions f, g uniformly on D prove that $(\lambda f_n + \mu g_n)$ converges to $\lambda f + \mu g$ uniformly on D for any $\lambda, \mu \in \mathbb{R}$.

Exercise 16.3 Which of the following formulae for $f_n(x)$ defines a sequence of functions (f_n) converging uniformly on $[0, 1]$?

(i) $x/(1+nx)$; (ii) nxe^{-nx^2}; (iii) $nx^n(1-x)$; (iv) $n^{1/2}x(1-x)^n$; (v) $nx(1-x^2)^{n^2}$; (vi) $x^n/(1+x^n)$; (vii) $n^{-x}x^n \cos nx$.

Exercise 16.4 Suppose that (f_n) converges to a function f uniformly on D. Prove that if each f_n is bounded on D then

(a) f is bounded on D;

(b) there is a uniform bound for the f_n, i.e. there exists $K \in \mathbb{R}$ such that $|f_n(x)| \leqslant K$ for all $x \in D$ and all $n \in \mathbb{N}$.

Exercise 16.5 Suppose that (f_n), (g_n) converge to f, g uniformly on D and that for each n the functions f_n, g_n are bounded on D. Prove that $(f_n g_n)$ converges to fg uniformly on D.

[Hint: use Exercise 16.4.]

Exercise 16.6 For each $n \in \mathbb{N}$ let $f_n(x) = 1/x$ and $g_n(x) = x/(1 + nx^2)$ for all $x \in (0, 1)$. Prove that (f_n) and (g_n) converge uniformly on $(0, 1)$ but $(f_n g_n)$ does not converge uniformly on $(0, 1)$.

Exercise 16.7 Construct functions $f_n : \mathbb{R} \to \mathbb{R}$ none of which is continuous at 0 but such that (f_n) converges uniformly on $\mathbb{R}$ to a continuous function.

Exercise 16.8 Suppose that (f_n) converges uniformly on D to a function f and that f_n is uniformly continuous on D for each n (recall Definition 13.23). Prove that f is uniformly continuous on D.

Exercise 16.9* (Dini's Theorem) Suppose that $f_n : X \to \mathbb{R}$ is a continuous function on a compact topological space X and that $f_n(x) \geqslant f_{n+1}(x)$ for all $n \in \mathbb{N}$ and all $x \in X$. Suppose also that (f_n) converges pointwise to a continuous function f on X. Prove that (f_n) converges uniformly on X.

17 Complete metric spaces

We saw in Chapter 4 how useful the completeness property of $\mathbb{R}$ is. From a theoretical viewpoint, completeness lets us solve equations such as $x^2 = 2$ in $\mathbb{R}$ which have no solution in $\mathbb{Q}$. Here is a practical version of the same phenomenon; we shall refer back to it a couple of times in this chapter.

Example 17.1 Let the sequence (s_n) be defined recursively as follows.

$$\text{Put } s_1 = 2 \text{ and, for any integer } n \geqslant 1, \text{ put } s_{n+1} = \frac{1}{2}\left(s_n + \frac{2}{s_n}\right).$$

Then (s_n) converges to $\sqrt{2}$.

Proof First we show inductively that $1 \leqslant s_n \leqslant 2$ for all positive integers n. This holds for $n = 1$, and if we assume $1 \leqslant s_n \leqslant 2$ then $1 \leqslant 2/s_n \leqslant 2$ so from the definition of s_{n+1} and the inductive hypothesis $1 \leqslant s_{n+1} \leqslant 2$.

Next we show that $s_n^2 \geqslant 2$ for all integers $n \geqslant 1$. This certainly holds for $n = 1$. Now $(s_n - 2/s_n)^2 \geqslant 0$ gives $s_n^2 + 4/s_n^2 \geqslant 4$, so for $n \geqslant 1$

$$s_{n+1}^2 = \frac{1}{4}\left(s_n + \frac{2}{s_n}\right)^2 = \frac{1}{4}\left(s_n^2 + \frac{4}{s_n^2} + 4\right) \geqslant 2.$$

We can now show that (s_n) is monotonic decreasing. For

$$\frac{s_{n+1}}{s_n} = \frac{1}{2}\left(1 + \frac{2}{s_n^2}\right) \leqslant 1 \quad \text{using } s_n^2 \geqslant 2.$$

This is the stage at which we need completeness: the sequence (s_n) is monotonic decreasing, and bounded below by 1, so by Proposition 4.16 (whose proof relies heavily on the completeness property of $\mathbb{R}$) (s_n) converges to a limit, say s. Now the algebra of limits applied to the formula defining s_{n+1} in terms of s_n gives $2s = s + 2/s$, so $s = \sqrt{2}$. □

Remark Readers who know some numerical analysis will recognise in Example 17.1 Newton's method for finding $\sqrt{2}$ more precisely, for finding successive approximations to $\sqrt{2}$. The method is fairly fast, enjoying what is called quadratic convergence; the first four terms of (s_n) are 2, 3/2, 17/12, 577/408. The fourth term already agrees with $\sqrt{2}$ to four decimal places.

In more general contexts it is often desirable to have an analogous completeness property, for example in order to guarantee existence of solutions to certain problems. The completeness property in Chapter 4 is convenient for studying continuity of real-valued functions of a real variable, but it uses the order properties of $\mathbb{R}$, and these do not generalize to metric spaces. However Cauchy's criterion for convergence of real number sequences is closely associated with Proposition 4.4, and it makes sense in any metric space. If it is true in a metric space X –that is if every Cauchy sequence in X converges—then we call X *complete.* In this chapter we study completeness of metric spaces We then study some applications via a particular result that uses completeness, Banach's fixed-point theorem.

Definition and examples

Definition 17.2 *A metric space X is* complete *if every Cauchy sequence in X converges (to a point of X).*

Example 17.3 (a) $\mathbb{R}$ is complete by Theorem 4.18.

(b) $\mathbb{Q}$ is not complete, for any sequence in $\mathbb{Q}$ which converges in $\mathbb{R}$ to an irrational number such as $\sqrt{2}$ is a Cauchy sequence in $\mathbb{Q}$ which does not converge to any point in $\mathbb{Q}$.

(c) $(0, 1) \subseteq \mathbb{R}$ is not complete, for the sequence $(1/n)$ is a Cauchy sequence in $(0, 1)$ which does not converge to any point in $(0, 1)$.

More examples of complete metric spaces, arising from the previous chapter, will be given shortly.

Example 17.3(a) and (c) show that completeness is not a topological property, since $(0, 1)$ and $\mathbb{R}$ are homeomorphic. However, it is invariant under uniform equivalence, in the sense of the next proposition.

Proposition 17.4 *Suppose that X, Y are metric spaces and there exists a bijective map $f : X \to Y$ such that both f and f^{-1} are uniformly continuous. Then X is complete iff Y is.*

The proof is contained in Exercise 17.6.

Corollary 17.5 *If metrics d_1, d_2 on a set X are Lipschitz equivalent (in the sense of Definition 6.33) then (X, d_1) is complete iff (X, d_2) is complete.*

Proof This follows since the identity map of X is uniformly continuous in both directions (see Proposition 13.25). □

The secret of Example 17.3(b), (c) is that these are not closed in $\mathbb{R}$.

Proposition 17.6 *A complete subspace Y of a metric space X is closed in X.*

Proof Suppose $x \in \overline{Y}$. By Exercise 6.26 there is a sequence (y_n) in Y converging to x in X. Since (y_n) is convergent it is Cauchy. So (y_n) is a Cauchy sequence in Y and by completeness of Y it must converge to a point in Y. By uniqueness of limits this means $x \in Y$, so Y is closed in X. □

The converse is also true.

Proposition 17.7 *A closed subspace Y of a complete metric space X is complete.*

Proof Suppose that (y_n) is a Cauchy sequence in Y. By completeness of X, there is an $x \in X$ such that (y_n) converges to x. Then, by Corollary 6.30, x is in Y, so Y is complete. □

Example 17.8 The closed intervals $[a, b]$, $(-\infty, b]$, $[a, \infty)$ are complete.

Next we study a few more basic results about completeness.

Proposition 17.9 *Any compact metric space X is complete.*

Proof Let (x_n) be any Cauchy sequence in X. Recall from Chapter 14 that both our possible definitions of compactness for a metric space imply sequential compactness, so from Definition 14.7 there is a subsequence $(x_{n(r)})$ converging to a point $x \in X$. The proof will now be completed by the next lemma, which tells us that the whole sequence (x_n) converges to x. □

Lemma 17.10 *If a Cauchy sequence (x_n) in a metric space X has a subsequence converging to $x \in X$ then (x_n) converges to x.*

Proof Call the metric d and suppose that the subsequence $(x_{n(r)})$ converges to x. Let $\varepsilon > 0$. Since (x_n) is Cauchy, there exists an integer N such that $d(x_m, x_n) < \varepsilon$ for all $m, n \geqslant N$. Since $(x_{n(r)})$ converges to x, there is some integer R such that $d(x_{n(r)}, x) < \varepsilon$ for all $r \geqslant R$. Let $n \geqslant N$, and choose $r \geqslant R$ such that $n(r) \geqslant N$. Then

$$d(x_n, x) \leqslant d(x_n, x_{n(r)}) + d(x_{n(r)}, x) < 2\varepsilon.$$

Hence (x_n) converges to x as required. □

The converse of Proposition 17.9 is not true in general (e.g. $\mathbb{R}$ is complete but not compact).

Proposition 17.11 *The product of two metric spaces (X, d_X), (Y, d_Y) is complete iff (X, d_X) and (Y, d_Y) are complete.*

Proof By Corollary 17.5 we may use any one of d_1, d_2, d_∞ to prove this, since these are Lipschitz equivalent metrics. We choose d_∞, whose definition we now recall: for (x_1, y_1), $(x_2, y_2) \in X \times Y$

$$d_\infty((x_1, y_1), (x_2, y_2)) = \max\{d_X(x_1, x_2), d_Y(y_1, y_2)\}.$$

Suppose that (X, d_X) and (Y, d_Y) are complete, and $((x_n, y_n))$ is a Cauchy sequence in $(X \times Y, d_\infty)$. Then $d_X(x_m, x_n) \leqslant d_\infty((x_m, y_m), (x_n, y_n))$, from which it follows easily that (x_n) is Cauchy. Likewise (y_n) is Cauchy. By completeness of (X, d_X) and (Y, d_Y) then (x_n) converges to some $x \in X$ and (y_n) converges to some $y \in Y$. Now $((x_n, y_n))$ converges to (x, y). For given $\varepsilon > 0$ there is an integer N_1 such that $d_X(x_n, x) < \varepsilon$ for all $n \geqslant N_1$ and an integer N_2 such that $d_Y(y_n, y) < \varepsilon$ for all $n \geqslant N_2$. So for any $n \geqslant \max\{N_1, N_2\}$ we have

$$d_\infty((x_n, y_n), (x, y)) = \max\{d_X(x_n, x), d_Y(y_n, y)\} < \varepsilon.$$

Hence $((x_n, y_n))$ converges to (x, y).

Conversely, suppose $(X \times Y, d_\infty)$ is complete. Let (x_n) be a Cauchy sequence in (X, d_X) and let y be any point in Y. Then one can check that (x_n, y) is Cauchy in $(X \times Y, d_\infty)$, for $d_\infty((x_m, y), (x_n, y)) = d_X(x_m, x_n)$. So by completeness of $(X \times Y, d_\infty)$ the sequence $((x_n, y))$ converges to some point (x, y') in $X \times Y$. It is clear that then $y' = y$ and (x_n) converges to x. Hence (X, d_X) is complete. Similarly (Y, d_Y) is complete. □

Corollary 17.12 *The product of a finite number of metric spaces is complete iff all the factors are complete.*

Corollary 17.13 $\mathbb{R}^n$ *is complete for each* $n \in \mathbb{N}$.

There are two more results about completeness, due to Cantor and Baire, which are valuable for advanced applications: we consider them on the web site.

Now we consider more examples of complete metric spaces, among function spaces that featured in the previous chapter. (There are further examples, involving sequence spaces, on the web site.)

Example 17.14 The space $(\mathcal{B}(D, \mathbb{R}), d_\infty)$ of bounded real-valued functions on a domain $D \subseteq \mathbb{R}$, with the sup metric d_∞, is complete. We have already mentioned this in Remark (b) after Theorem 16.9.

As we shall see below, this is important for guaranteeing solutions to certain problems just as the completeness property of $\mathbb{R}$ was important for Example 17.1.

★ More generally we may consider the space $\mathcal{B}(D, X)$ of all bounded functions from a set D to a metric space X, with the sup metric d_∞ as before.

Proposition 17.15 *The space $(\mathcal{B}(D, X), d_\infty)$ is complete iff X is complete.*

Proof A close scrutiny of the proof of Theorem 16.9 reveals that it uses nothing about D, and only the completeness of $\mathbb{R}$, so the same proof shows that $(\mathcal{B}(D, X), d_\infty)$ is complete if X is complete. The converse is easy: If X is not complete, take any Cauchy sequence (x_n) in X which fails to converge in X, and the corresponding sequence of constant functions (with values x_n) will yield a non-convergent Cauchy sequence in $\mathcal{B}(D, X)$. ★ □

Example 17.16 Let $(\mathcal{B}(D, \mathbb{R}), d_\infty)$ be as in 17.14, where $D \subseteq \mathbb{R}$, and as usual we take the Euclidean topology on D. We shall show that the subspace $\mathcal{C}(D, \mathbb{R})$ of continuous bounded real-valued functions on D is complete. First let $\mathcal{C}_c$ be the subspace of those bounded functions which are also continuous at some given point $c \in D$. From Theorem 16.10 we deduce that $\mathcal{C}_c$ is a closed subspace of $\mathcal{B}(D, \mathbb{R})$ and hence by Proposition 17.7 it is complete. Now $\mathcal{C}(D, \mathbb{R})$ is the intersection of the family $\{\mathcal{C}_c : c \in D\}$ so it too is closed in $\mathcal{B}(D, \mathbb{R})$ and hence complete.

★ More generally, if X is any topological space and Y is any complete metric space then the space of all continuous bounded maps from X to Y, with the sup metric, is complete by a similar argument. ★

Example 17.17 The space (X, d_1) of continuous real-valued functions on $[0, 1]$ with the L^1 metric of Example 5.14 is not complete. We shall look at this in detail since it is a rich source of fallacies. To construct a Cauchy sequence which does not converge in $L^1[0, 1]$ we define $f_n : [0, 1] \to \mathbb{R}$

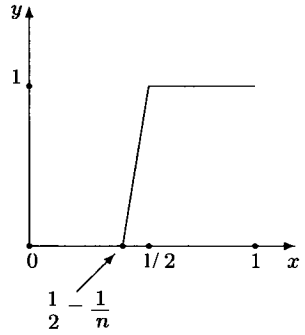

Figure 17.1. Graph of f_n

for $n \geqslant 2$ by

$$f_n(x) = \begin{cases} 0, & x \in \left[0, \frac{1}{2} - \frac{1}{n}\right], \\ 1, & x \in \left[\frac{1}{2}, 1\right], \\ n\left(x + \frac{1}{n} - \frac{1}{2}\right), & x \in \left[\frac{1}{2} - \frac{1}{n}, \frac{1}{2}\right], \end{cases}$$

Then (f_n) is a Cauchy sequence in $L^1[0, 1]$: for given $\varepsilon > 0$ let N be an integer with $N > 1/\varepsilon$ and $m > n \geqslant N$; then $f_m(x) = f_n(x)$ except in the interval $I_n = \left[\frac{1}{2} - \frac{1}{n}, \frac{1}{2}\right]$, and $|f_m(x) - f_n(x)| \leqslant 1$ for $x \in I_n$ (see Figure 17.1). We therefore get

$$d_1(f_m, f_n) = \int_0^1 |f_m - f_n| < 1/n < 1/N < \varepsilon \quad \text{for all} \quad m > n \geqslant N.$$

It remains to prove that (f_n) does not converge in (X, d_1). First here are two fallacious proofs.

(a) For $x \in [0, 1/2)$, $f_n(x) \to 0$ as $n \to \infty$, and for $x \in [1/2, 1]$, $f_n(x) \to 1$ as $n \to \infty$. Hence the limit function f is not continuous on $[0, 1]$, ($f(x)$ is zero on $[0, 1/2)$ and 1 on $[1/2, 1]$), is therefore not in X, so (f_n) does not converge in X.

What this proves is that (f_n) converges pointwise to a discontinuous function on $[0, 1]$. From this we can conclude only that (f_n) does not converge uniformly to f it is still conceivable that there is some continuous function to which it converges in the L^1 metric d_1. To clarify the fallacy here, suppose we let $g_n = f_n|[0, 1/2]$ and consider (g_n) in the space X' analogous to X but with $[0, 1]$ replaced by $[0, 1/2]$. Then the argument in (a) would say that (g_n) converges to the discontinuous function $f|[0, 1/2]$

hence does not converge in X'. However, in fact (g_n) does converge in (X', d_1), namely to the zero function g, since $\int_0^{1/2} |g_n - g| = 1/2n \to 0$ as $n \to \infty$.

(b) Let f be as in (a). Then $\int_0^1 |f_n - f| = 1/2n \to 0$ as $n \to \infty$. Hence (f_n) converges to f in the L^1 metric, and since limits of sequences in a metric space are unique, (f_n) cannot converge to a continuous function in the L^1 metric.

This is closer: at least it uses the right metric. But it will still not do as it stands. For f belongs to, say, the set $\mathcal{R}$ of Riemann-integrable functions on $[0, 1]$. (Note that $X \subseteq \mathcal{R}$.) Now the formula $d_1(f, g) = \int_0^1 |f - g|$ does not define a metric in this larger set, but only what is called a pseudo-metric, for the axiom (M1) in general fails: for example f and g might differ at only a finite set of points, and we would have $d_1(f, g) = 0$ but $f \neq g$.

(c) However, we can develop (b) into a valid proof. Let f be as before, so $d_1(f_n, f) \to 0$ as $n \to \infty$. Suppose that (f_n) does converge to a continuous function g in the L^1 metric, so $\int_0^1 |f_n - g| \to 0$ as $n \to \infty$. Since

$$|f(x) - g(x)| \leqslant |f(x) - f_n(x)| + |f_n(x) - g(x)| \quad \text{for all } x \in [0, 1],$$

it follows by integration that

$$0 \leqslant \int_0^1 |f - g| \leqslant \int_0^1 |f - f_n| + \int_0^1 |f_n - g| = \theta_n, \text{ say.}$$

Since $\theta_n \to 0$ as $n \to \infty$ and $\int_0^1 |f - g|$ is independent of n we deduce that this latter integral is zero. By the proof of Lemma 5.15, $f(x) = g(x)$ at any $x \in [0, 1]$ at which $f - g$ is continuous. Since g is continuous on $[0, 1]$ by hypothesis and f is continuous on $[0, 1]$ except at $1/2$, it follows that $g(x) = f(x)$ for $x \neq 1/2$, which is clearly impossible for a continuous function g. Hence we finally have a proof that (X, d_1) is not complete.

If a metric space X is not complete, we can construct a complete metric space $\hat{X}$, called the completion of X, such that X is (isometric to) a dense subspace of $\hat{X}$; this is like extending from $\mathbb{Q}$ to $\mathbb{R}$. The general construction is described on the web site.

Banach's fixed point theorem

Our applications of completeness will be made via this theorem, which is one of metric space theory's most attractive results. On the theoretical side, it unifies the proofs of several existence theorems for solutions of algebraic, differential, integral and functional equations. The theorem, when it works, tells us that the solution is unique as well. It employs a constructive method which is of also interest in numerical analysis.

Definition 17.18 *Given any self-map $f : S \to S$ of a set S, a* fixed point of f *is a point $p \in S$ such that $f(p) = p$.*

For example, a rotation of a disc around its centre has the centre as a fixed point. The conscientious reader will already have proved in Exercise 12.8 that any continuous self-map of $[a, b]$ has at least one fixed point. We begin with a special case of Banach's theorem which proves, under stronger hypotheses than continuity, that there is a unique fixed point. We need some terminology. Let $D \subseteq \mathbb{R}$.

Definition 17.19 *For given positive real numbers α and K, a function $f : D \to \mathbb{R}$* satisfies a Lipschitz condition of order α on D, with constant K, *if*

$$|f(x) - f(y)| \leqslant K|x - y|^{\alpha} \quad \text{for all} \quad x, y \in D.$$

The next proposition, whose proof is Exercise 17.7, is intended to relate Lipschitz conditions to familar properties.

Proposition 17.20 *(a) If f satisfies a Lipschitz condition of order $\alpha > 0$ on D then f is uniformly continuous on D.*

(b) If f satisfies a Lipschitz condition of order $\alpha > 1$ on $[a, b]$ then f is constant on $[a, b]$.

(c) If $f : [a, b] \to \mathbb{R}$ is continuous on $[a, b]$ and differentiable on (a, b) with $|f'(x)| \leqslant K$ for all $x \in (a, b)$ then f satisfies a Lipschitz condition of order 1 with constant K on $[a, b]$.

It is (c) that often allows us to deduce that a given function satisfies a Lipschitz condition.

Example 17.21 Consider the real-valued function of a real variable with formula

$$f(x) = \frac{1}{2}\left(x + \frac{2}{x}\right).$$

Then f defines a function $f : [1, 2] \to [1, 2]$, and this function satisfies a Lipschitz condition of order 1 with constant $1/2$ on $[1, 2]$.

Proof This is closely connected to Example 17.1, in which $s_{n+1} = f(s_n)$. The proof that f maps $[1, 2]$ into itself is the same as the proof that $1 \leqslant s_n \leqslant 2$ implies $1 \leqslant s_{n+1} \leqslant 2$. To get a Lipschitz condition, note that

$$f'(x) = \frac{1}{2}\left(1 - \frac{2}{x^2}\right), \quad \text{so } |f'(x)| \leqslant \frac{1}{2} \quad \text{for all } x \in [1, 2].$$

The result now follows by Proposition 17.20(c). □

Remark If we wanted a smaller K, for example, to use in Proposition 17.27 below, we could prove that $f(x) \geqslant \sqrt{2}$ for all $x \geqslant \sqrt{2}$ and that $f(x) \leqslant 3/2$ for all $x \in [\sqrt{2}, 3/2]$. Thus f maps $[\sqrt{2}, 3/2]$ into itself, and on this interval we may take $K = 1/18$.

Theorem 17.22 *If $f : [a, b] \to [a, b]$ satisfies a Lipschitz condition of order 1 with constant $K < 1$ on $[a, b]$ then f has a unique fixed point p in $[a, b]$. Moreover, if x_1 is any point in $[a, b]$ and $x_n = f(x_{n-1})$ for $n > 1$, then (x_n) converges to p. The same result holds if $[a, b]$ is replaced throughout by $(-\infty, b]$ or $[a, \infty)$.*

This is a special case of Banach's fixed point theorem, which will be proved in general later. Figure 17.2 illustrates two possible cases.

Example 17.23 The function $f : [1, 2] \to [1, 2]$ defined by

$$f(x) = \frac{1}{2}\left(x + \frac{2}{x}\right) \quad \text{has a unique fixed point } p \text{ in } [1, 2].$$

This fixed point p is the limit s of the sequence in Example 17.1. It satisfies $2p = (p + 2/p)$ so $p = \sqrt{2}$.

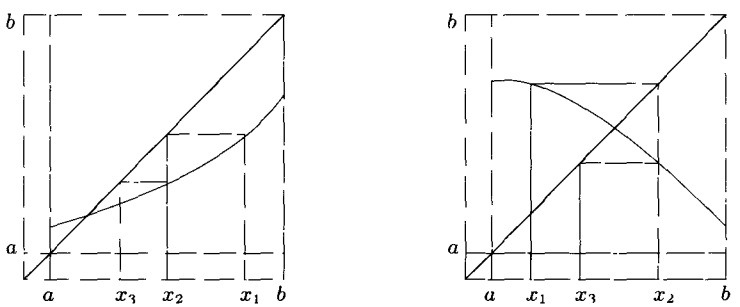

Figure 17.2. Banach's theorem for $[a, b]$

Contraction mappings

Lipschitz conditions easily generalize to metric spaces. The kind of condition we are concerned with has an appealing geometric interpretation.

Definition 17.24 *Let (X, d) be a metric space. A map $f : X \to X$ is a* contraction *if for some constant $K < 1$ we have $d(f(x), f(y)) \leqslant Kd(x, y)$ for all x, y in X.*

Lemma 17.25 *Any contraction of a metric space X is uniformly continuous.*

Proof With the above notation, let $\varepsilon > 0$ and take $\delta = \varepsilon/K$. For any $x, y \in X$ satisfying $d(x, y) \leqslant \delta$ we have $d(f(x), f(y)) \leqslant Kd(x, y) < \varepsilon$ as required. □

Theorem 17.26 (Banach's fixed point theorem) *If $f : X \to X$ is a contraction of a complete metric space X then f has a unique fixed point p in X.*

Proof *Existence* Let x_1 be chosen arbitrarily in X and inductively let $x_n = f(x_{n-1})$ for $n > 1$. We shall prove that (x_n) is a Cauchy sequence.

First we note that by an easy induction $d(x_{r+1}, x_r) \leqslant K^{r-1}d(x_2, x_1)$ for all $r > 1$. Now for $m > n$, by repeated use of the triangle inequality,

$$d(x_m, x_n) \leqslant d(x_m, x_{m-1}) + d(x_{m-1}, x_{m-2}) + \ldots + d(x_{n+1}, x_n).$$

Hence

$$\begin{aligned} d(x_m, x_n) &\leqslant (K^{m-2} + K^{m-3} + \ldots + K^{n-1})d(x_2, x_1) \\ &= K^{n-1}(K^{m-n-1} + K^{m-n-2} + \ldots + 1)d(x_2, x_1) \\ &= \frac{K^{n-1}(1 - K^{m-n})}{1 - K}d(x_2, x_1) \\ &< \frac{K^{n-1}}{1 - K}d(x_2, x_1). \end{aligned}$$

Now $K^{n-1} \to 0$ as $n \to \infty$, since $0 \leqslant K < 1$. Hence for any $\varepsilon > 0$ we have $d(x_m, x_n) < \varepsilon$ whenever $m \geqslant n$ and n is sufficiently large (specifically, when n is such that $\dfrac{K^{n-1}}{1 - K}d(x_2, x_1) < \varepsilon$). This proves that (x_n) is a Cauchy sequence. Since X is complete, (x_n) converges to some point $p \in X$. By continuity of f at p (Lemma 17.25), $f(x_n) \to f(p)$ as $n \to \infty$. But $f(x_n) = x_{n+1} \to p$ as $n \to \infty$, so $f(p) = p$.

Uniqueness If $f(p) = p$ and also $f(q) = q$ then

$$d(p, q) = d(f(p), f(q)) \leqslant Kd(p, q).$$

Since $K < 1$ this is a contradiction unless $d(p, q) = 0$. Hence $q = p$ and the fixed point is unique. □

From a practical point of view, the next result is also of interest.

Proposition 17.27 *With the notation and hypotheses of Theorem 17.26,*

$$d(p, x_n) \leqslant \frac{K^{n-1}}{1-K} d(x_2, x_1).$$

Proof As above, for $m > n$ we have $d(x_m, x_n) \leqslant \dfrac{K^{n-1}}{1-K} d(x_2, x_1)$. Keep n fixed and let $m \to \infty$. Then $x_m \to p$, and by continuity of d (see Exercise 5.17) $d(x_m, x_n) \to d(p, x_n)$ as $m \to \infty$. In the limit we get the result. □

In this error estimate of how far x_n is from p, the right-hand side can be calculated in specific cases without knowing p in advance.

Remarks (a) Banach's fixed point theorem is also known as the contraction mapping theorem.

(b) There are many variations on the contraction mapping theme. For example, it is enough to assume that some iterate $f^{(n)}$ of f, rather than f itself, is a contraction (see Exercise 17.13). However, as Exercises 17.11 and 17.12 illustrate, we cannot simply drop completeness or the uniform factor K.

(c) The statement that f is a contraction depends on the choice of metric d. It is possible that d and d' are uniformly equivalent metrics on X, in which case (X, d) is complete iff (X, d') is complete, and that $f : X \to X$ is a d-contraction but not a d'-contraction. This gives scope for ingenuity in the choice of metric in applications, see for example the proof of Theorem 17.31 below.

Applications of Banach's fixed point theorem

We end with some applications of Banach's fixed point theorem, in addition to Theorem 17.22. There is a further application on the web site —an inverse function theorem for functions of several real variables.

Example 17.28 First consider the (possibly unfamiliar) equation

$$\psi(x) = \lambda \int_a^b K(x, y)\psi(y)\,dy + f(x),$$

where K and f are known, and the problem is to find ψ. This is called an integral equation. Integral equations help in the qualitative study of differential equations, as we shall illustrate below. Also, certain applied mathematics problems naturally give rise to integral equations.

Suppose that in the above, $K : [a, b] \times [a, b] \to \mathbb{R}$ and $f : [a, b] \to \mathbb{R}$ are continuous and that $\lambda \in \mathbb{R}$. We shall prove that if $|\lambda|$ is sufficiently small then there is a unique continuous function $\psi : [a, b] \to \mathbb{R}$ satisfying the integral equation.

Let $X = \mathcal{C}[a, b]$ be the space of all continuous real-valued functions on $[a, b]$ with the sup metric d_∞ as in Example 5.13. By Example 17.16 X is complete. Define $F : X \to X$ as follows: for any function $\psi \in X$, let $F(\psi) : [a, b] \to \mathbb{R}$ be given by

$$F(\psi)(x) = \lambda \int_a^b K(x, y)\psi(y)\,dy + f(x),$$

for all $x \in [a, b]$. It follows easily from integration theory that $F(\psi)$ is continuous on $[a, b]$, so $F(\psi) \in X$. We shall prove that for $|\lambda|$ sufficiently small, F is a contraction.

Since $[a, b] \times [a, b]$ is compact and K is continuous, there exists Δ such that $|K(x, y)| \leqslant \Delta$ for all x, y in $[a, b]$. For any ψ_1, $\psi_2 \in X$,

$$\begin{aligned} d_\infty(F(\psi_1), F(\psi_2)) &= \sup_{x \in [a, b]} |F(\psi_1)(x) - F(\psi_2)(x)| \\ &= \sup_{x \in [a, b]} \left| \lambda \int_a^b K(x, y)(\psi_1(y) - \psi_2(y))\,dy \right| \\ &\leqslant |\lambda|(b - a)\Delta \sup_{y \in [a, b]} |\psi_1(y) - \psi_2(y)|, \end{aligned}$$

where the last inequality follows from integration theory.

Since $\sup_{x \in [a, b]} |\psi_1(y) - \psi_2(y)| = d_\infty(\psi_1, \psi_2)$, it follows that F is a contraction provided $|\lambda| < \{(b - a)\Delta\}^{-1}$. By Theorem 17.26, F has a unique fixed point in X for such λ. But $F(\psi) = \psi$ means that ψ is a solution to the original integral equation, so the integral equation has a unique continuous solution for such λ.

Theorem 17.29 uses Remark (b) above, whose proof is Exercise 17.13.

Theorem 17.29 *Suppose that* $K : [a, b] \times [a, b] \to \mathbb{R}$ *and* $f : [a, b] \to \mathbb{R}$ *are continuous. Then the (Volterra) equation*

$$\phi(x) = f(x) + \int_a^x K(x, y)\phi(y)\,\mathrm{d}y$$

has a unique continuous solution ϕ *on* $[a, b]$.

Proof Let T be the triangular set $\{(x, y) : a \leqslant y \leqslant x \leqslant b\}$. Since T is bounded and closed in $\mathbb{R}^2$, it is compact. Let Δ be an upper bound for $|K(x, y)|$ on T. Let X be the complete metric space of continuous real-valued functions on $[a, b]$ with the sup metric d_∞. Define $F : X \to X$ by

$$F(\phi)(x) = f(x) + \int_a^x K(x, y)\phi(y)\,\mathrm{d}y \quad \text{for all } x \in [a, b].$$

Then
$$\begin{aligned} d_\infty(F(\phi_1), F(\phi_2)) &= \sup_{x\in[a,b]} \left| \int_a^x K(x, y)(\phi_1(y) - \phi_2(y))\,\mathrm{d}y \right| \\ &\leqslant (b-a)\Delta \sup_{y\in[a,b]} |\phi_1(y) - \phi_2(y)| \\ &= (b-a)\Delta d_\infty(\phi_1, \phi_2). \end{aligned}$$

This is not enough to make F a contraction, unless $(b-a)\Delta < 1$. But we can now use Exercise 17.13: to get a unique fixed point for F it is enough to show that some iterate $F^{(n)}$ of F is a contraction. Inductively suppose that

$$|F^{(r)}(\phi_1)(x) - F^{(r)}(\phi_2)(x)| \leqslant \frac{(x-a)^r}{r!}\Delta^r d_\infty(\phi_1, \phi_2) \quad \text{for all } x \in [a, b].$$

This certainly holds for $r = 0$, taking $F^{(0)}$ to mean the identity function. Then for any $x \in [a, b]$,

$$\begin{aligned} &|F^{(r+1)}(\phi_1)(x) - F^{(r+1)}(\phi_2)(x)| \\ &\quad = \left| \int_a^x K(x, y)[F^{(r)}(\phi_1)(y) - F^{(r)}(\phi_2)(y)]\,\mathrm{d}y \right| \\ &\quad \leqslant \Delta \left| \int_a^x \frac{(y-a)^r}{r!}\Delta^r d_\infty(\phi_1, \phi_2)\,\mathrm{d}y \right| \\ &\quad \leqslant \frac{|x-a|^{r+1}}{(r+1)!}\Delta^{r+1} d_\infty(\phi_1, \phi_2). \end{aligned}$$

Hence in particular

$$d_\infty(F^{(n)}(\phi_1), F^{(n)}(\phi_2)) \leqslant \frac{(b-a)^n\Delta^n}{n!} d_\infty(\phi_1, \phi_2) \text{ for all integers } n \geqslant 0$$

and all ϕ_1, $\phi_2 \in X$. Since $(b-a)^n\Delta^n/n! \to 0$ as $n \to \infty$, we see that the iterate $F^{(n)}$ is a contraction for n sufficiently large. Hence by Exercise 17.13, F has a unique fixed point in X, which tells us that the original Volterra integral equation has a unique continuous solution on $[a, b]$. □

Example 17.30 We end with the Cauchy-Picard theorem of differential equations, which says that under suitable conditions on f, the initial value problem

$$\frac{dy}{dx} = f(x, y); \qquad y(x_0) = y_0$$

has a unique solution in a small enough neighbourhood of (x_0, y_0). More precisely:

Theorem 17.31 *Suppose that* $f : D = [x_0-a, x_0+a]\times[y_0-b, y_0+b] \to \mathbb{R}$ *is continuous and satisfies the 'Lipschitz condition'*

$$|f(x, y_1)-f(x, y_2)| \leqslant K|y_1-y_2| \text{ for all } (x, y_1), (x, y_2) \in D, \text{ some } K > 0.$$

Let M *be an upper bound for* $|f(x, y)|$ *on* D, *and let* $c = \min\{a, b/M\}$. *Then on* $I = [x_0 - c, x_0 + c]$ *there exists a unique solution* y *of the differential equation* $\dfrac{dy}{dx} = f(x, y)$ *such that* $y(x_0) = y_0$.

Proof The traditional proof uses a method of successive approximations and properties of uniform convergence, which are codified in Banach's fixed point theorem. We first prove a weaker version which uses the sup metric, and then show how fiddling with the metric gives a proof of the full theorem.

We need to use a few facts from calculus. First we see that a function y is a solution of the initial value problem on I iff y is a continuous solution of the integral equation

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))\,dt \quad \text{on } I.$$

For suppose y is a continuous solution of this integral equation. Since y and f are continuous, the integrand is continuous on I; remembering how to differentiate such an integral we see that $dy/dx = f(x, y)$, and putting $x = x_0$ we see that $y(x_0) = y_0$. Conversely if $y(x)$ satisfies the initial value problem on I then integration over the interval from x_0 to x tells us that

$$\int_{x_0}^{x} y'(t)\,dt = \int_{x_0}^{x} f(t, y(t))\,dt \quad \text{for any } x \in I.$$

By the Fundamental theorem of calculus, the left-hand side of this equation is $y(x) - y(x_0)$, so $y(x)$ satisfies the integral equation, and $y(x)$ is continuous on I by integration theory.

Now we apply Banach's fixed point theorem to the integral equation. We first prove a weak version of the Cauchy-Picard theorem, in which I is replaced by the (probably shorter) interval $I' = [x_0 - c', x_0 + c']$ where $c' < \min\{a, b/M, 1/K\}$. Let and let $X' \subset \mathcal{C}I'$ be the space of continuous real-valued functions y on I' such that $|y(x) - y_0| \leqslant b$ for all $x \in I'$. Then X' is closed in $\mathcal{C}I'$, since if $|y(x) - y_0| > b$ for some $x \in I'$ then the same will be true for all functions in $\mathcal{C}I'$ close enough to $y(x)$ in the sup metric, so the complement of X' in $\mathcal{C}I'$ is open in $\mathcal{C}I'$. Hence X' is complete, since $\mathcal{C}I'$ is complete by Example 17.16. Now define $F : X' \to X'$ by

$$F(y)(x) = y_0 + \int_{x_0}^{x} f((t,\ y(t))\,\mathrm{d}t.$$

First, $F(y) \in X'$ when $y \in X'$; for $F(y)$ is continuous on I' by integration theory; also,

$$|F(y)(x) - y_0| = \left| \int_{x_0}^{x} f((t,\ y(t))\,\mathrm{d}t \right| \leqslant c'M \leqslant b$$

for any $x \in I'$. Next, F is a contraction, for if $y_1,\ y_2 \in X'$ then

$$d_\infty(y_1,\ y_2) = \sup_{|x - x_0| \leqslant c'} \left| \int_{x_0}^{x} \{f(t,\ y_1(t)) - f(t,\ y_2(t))\}\,\mathrm{d}t \right|$$

$$\leqslant c'K d_\infty(y_1, y_2), \quad \text{and} \quad c'K < 1.$$

Hence F has a unique fixed point, which is a unique continuous solution of the original initial value problem on I'.

However, we can establish existence and uniqueness over all of I by using a different metric on $\mathcal{C}I$ (see Remark (c) above). For $y_1,\ y_2 \in \mathcal{C}I$ define

$$d(y_1,\ y_2) = \sup_{x \in I} \mathrm{e}^{-K|x - x_0|} |y_1(x) - y_2(x)|.$$

Note that for any $t \in I$ we have

$$\mathrm{e}^{-K|t - x_0|} |y_1(t) - y_2(t)| \leqslant \sup_{x \in I} \mathrm{e}^{-K|x - x_0|} |y_1(x) - y_2(x)| = d(y_1,\ y_2),$$

so for any $t \in I$

$$|y_1(t) - y_2(t)| \leqslant d(y_1,\ y_2)\mathrm{e}^{K|t - x_0|}.$$

The metric d is Lipschitz equivalent (see Definition 6.33) to the sup metric on $\mathcal{C}I$, since $\mathrm{e}^{-Kc} \leqslant \mathrm{e}^{-K|x - x_0|} \leqslant 1$ for all $x \in I$. Hence $\mathcal{C}I$ with this metric

is complete (see Exercise 17.4). Let $X \subseteq \mathcal{C}I$ consist of the y in $\mathcal{C}I$ satisfying $|y(x) - y_0| \leqslant b$ for all $x \in I$. This is a closed subspace of $\mathcal{C}I$ just as before, so is also complete. Define $F : X \to X$ as before. We can check that $F(y) \in X$ when $y \in X$. But now for $y_1, y_2 \in X$,

$$\begin{aligned} d(F(y_1), F(y_2)) &= \sup_{x \in I} \mathrm{e}^{-K|x-x_0|} \left| \int_{x_0}^{x} \{f(t, y_1(t)) - f(t, y_2(t))\} \, \mathrm{d}t \right| \\ &\leqslant \sup_{x \in I} \mathrm{e}^{-K|x-x_0|} \left| \int_{x_0}^{x} K|y_1(t) - y_2(t)| \, \mathrm{d}t \right| \\ &\leqslant d(y_1, y_2) \sup_{x \in I} \mathrm{e}^{-K|x-x_0|} \left| \int_{x_0}^{x} K\mathrm{e}^{K|t-x_0|} \, \mathrm{d}t \right| \\ &= d(y_1, y_2) \sup_{x \in I} \mathrm{e}^{-K|x-x_0|} (\mathrm{e}^{K|x-x_0|} - 1) \\ &= d(y_1, y_2) \sup_{x \in I} (1 - \mathrm{e}^{-K|x-x_0|}) \\ &\leqslant (1 - \mathrm{e}^{-Kc}) d(y_1, y_2). \end{aligned}$$

Hence F is a contraction, since $1 - \mathrm{e}^{-Kc} < 1$. This shows that the original initial value problem has a unique solution over the interval I, thereby establishing the full force of Picard's Theorem. □

Exercise 17.1 Prove that any discrete metric space is complete.

Exercise 17.2 Which of the following are complete?

(i) $\{1/n \,.\, n \in \mathbb{N}\} \cup \{0\}$, (ii) $\mathbb{Q} \cap [0, 1]$. (iii) $\{(x, y) \in \mathbb{R}^2 : x > 0, y \geqslant 1/x\}$.

Exercise 17.3 Prove that in any metric space

(a) the union of two complete subspaces is complete,

(b) the intersection of a family of complete subspaces is complete (provided it is non-empty).

[Hint: for (a) you could use Lemma 17.10.]

Exercise 17.4 Prove that if metrics d, d' on a set X are Lipschitz equivalent (see Definition 6.33) and (X, d) is complete then so is (X, d').

Exercise 17.5★ Recall the following metrics for $\mathbb{R}$ from Exercise 5.4·

$(a)\, d(x, y) = |x^3 - y^3|$, $(b)\, d(x, y) = |\mathrm{e}^x - \mathrm{e}^y|$, $(c)\, d(x, y) = |\tan^{-1}(x) - \tan^{-1}(y)|$.

For which of these metrics is $(\mathbb{R}, d)$ is complete?

Exercise 17.6* Suppose that $f : X \to X'$ is a one-one correspondence of metric spaces such that f is uniformly continuous and f^{-1} is continuous.

(a) Prove that if (x_n) is a Cauchy sequence in X then $f(x_n)$ is a Cauchy sequence in X'.

(b) Prove that if (y_n) is a convergent sequence in X' then $(f^{-1}(y_n))$ is a convergent sequence in X.

(c) Deduce that if X' is complete then X is complete.

(d) Deduce that completeness is an invariant of uniform equivalence (recall that a uniform equivalence is a bijective map which together with its inverse is uniformly continuous).

Exercise 17.7 Prove Proposition 17.20, that

(a) If f satisfies a Lipschitz condition of order $\alpha > 0$ on D then f is uniformly continuous on D.

(b) If f satisfies a Lipschitz condition of order $\alpha > 1$ on $[a, b]$ then f is constant on $[a, b]$.

(c) If f is continuous on $[a, b]$ and differentiable on (a, b) with $|f'(x)| \leqslant K$ for all $x \in (a, b)$ then f satisfies a Lipschitz condition of order 1 with constant K on $[a, b]$.

[Hint: use the mean value theorem for (b) and (c).]

Exercise 17.8 Use Theorem 17.22 to prove that the equation $x^5 + 7x - 1 = 0$ has a unique solution in $[0, 1]$.

Exercise 17.9 Show that the cosine function $\cos : [0, 1] \to [0, 1]$ is a contraction. Find an approximate solution to the equation $\cos x = x$, correct to two decimal places.

Exercise 17.10 Let $x_1 = \sqrt{2}$ and for $n \geqslant 1$ let $x_{n+1} = \sqrt{2 + \sqrt{x_n}}$. Use Banach's fixed point theorem to show that (x_n) converges to a root of the equation $x^4 - 4x^2 - x + 4 = 0$ lying between $\sqrt{3}$ and 2.

Exercise 17.11 Define $f : (0, 1/4) \to (0, 1/4)$ by $f(x) = x^2$. Show that f is a contraction which has no fixed point.

Exercise 17.12 Define $f\ .\ [1, \infty) \to [1, \infty)$ by $f(x) = x + x^{-1}$. Show that $[1, \infty)$ is complete and $|f(x) - f(y)| < |x - y|$ for any distinct $x, y \in [1, \infty)$, yet f has no fixed point.

Exercise 17.13 (a) Suppose that $f : X \to X$ is a map of a complete metric space X and for some integer k, the iterated map $f^{(k)} = f \circ f \circ \ldots \circ f$ (k times) is a contraction. Prove that f has a unique fixed point p in X and * that for any $x \in X$ the sequence $(f^{(n)}(x))$ converges to $p^\star$.

(b) Show that the cosine function $\cos : \mathbb{R} \to \mathbb{R}$ is not a contraction but that $\cos^{(2)}$ is a contraction.

Exercise 17.14 Let $X = \{x, y, z\}$, let d be the discrete metric on X and let

$$d'(x, y) = 2,\ d'(y, x) = 2,\ d'(x, z) = 2,\ d'(z, x) = 2,\ d'(y, z) = 1,\ d'(z, y) = 1$$

$$d'(x, x) = d'(y, y) = d'(z, z) = 0.$$

(a) Show that d' is a metric for X which is Lipschitz equivalent to d.

(b) Define $f : X \to X$ by $f(x) = y$, $f(y) = z$, $f(z) = z$. Prove that f is a d'-contraction but not a d-contraction.

Exercise 17.15* Let $f : X \to X$ be a map of a compact metric space X such that $d(f(x), f(y)) < d(x, y)$ for any distinct points $x, y \in X$. Prove that f has a unique fixed point.
[Hint: show that $\inf\{d(x, f(x)) : x \in X\}$ is attained, and get a contradiction unless this inf is zero.]

Exercise 17.16 Let $\mathcal{C}[0, 1]$ denote the complete metric space of continuous real-valued functions on the closed interval $[0, 1]$ equipped with the sup metric.

(a) Show that the function $I : \mathcal{C}[0, 1] \to \mathbb{R}$ defined by $I(f) = \int_0^1 f(x)\,\mathrm{d}x$ is continuous.

(b) Let $g \in \mathcal{C}[0, 1]$ and let $F : \mathcal{C}[0, 1] \to \mathcal{C}[0, 1]$ be defined by

$$F(y)(x) = g(x) + \frac{1}{2}\int_0^1 \sin(xt)y(t)\,\mathrm{d}t.$$

Show that F is a contraction mapping and deduce that the equation $y = F(y)$ has a unique solution in $\mathcal{C}[0, 1]$.

Bibliography

Hart, M. (2001) *Guide to Analysis (Macmillan Mathematical Guides* (2nd rev edn)) Palgrave Macmillan

Hewitt, E.(1960) 'The rôle of compactness in analysis' *Amer Math Monthly,* **67** 499–516

Lawson, Terry (2003) *Topology· A Geometric Approach.* Oxford University Press

Mendelson, B (1990). *Introduction to topology* Dover Publications

Munkres, James R (2000) *Topology* (2nd edn) Prentice Hall

Priestley, H A (2003) *Introduction to Complex Analysis* (2nd edn) Oxford University Press.

Spivak, M (2006) *Calculus* (3rd edn). Cambridge University Press

Willard, S (2004) *General Topology.* Dover Publications

Index

Bold type indicates principal references (for example, definitions)